SIMULATION TECHNIQUES TOWARDS RECOGNIZING TAMIL CHARACTERS INSCRIBED IN ANTIQUATED COPPER-PLATES: APPLICATION OF COMPLEX DEEP-LEARNING MACHINE ON ECO- FRIENDLY CLEANSED PLATES

A Thesis

Submitted by

R. INDRA GANDHI

[Reg. No. D15CS512]

In partial fulfilment for the award of the degree

of

DOCTOR OF PHILOSOPHY

Under the Supervision of

Dr. M. PONNAVAIKKO

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF SCIENCE AND HU ENGINEERING

BHARATH INSTITUTE OF HIGHER EDUCATION AND RESEARCH

CHENNAI -600 073

JANUARY 2021

CHAPTER 9

9. CONCLUSION AND FUTURE SCOPE

9.1 CONCLUSION

In this chapter, the main contributions delivered and the significant achievement acquired from this research work is summarized. The conclusion, which follows the summary, highlights the research contributions delivered in the field of copper plate Tamil character recognition using image processing. Moreover, on the view of providing the future exploring possibilities to researches that follow, the present limitations and expansion possibilities of this system are also briefed.

Copper Plate Optical Character Recognition (CPOCR) for composed content is at present an open territory of research. The main aspect of this thesis is to concentrate on Tamil character recognition using copper plate Tamil text image restoration process. The present constraints in these fields of character recognition are either related to feature extraction or classification difficulties. This thesis is focused on overcoming these difficulties faced on feature extraction and classification as both of them have equally important roles to play in character recognition.

Tamil scripts are normally grouped into four classes namely Vowels, Consonants, Composite characters and Aydham. These four classes are taken for classification purpose in this research work. Traditional algorithms are far slower than required

because of their gradient based learning algorithm and the parameters have to be tuned iteratively. And therefore, Extreme Learning Machine (ELM) is used for classification.

The performance of ELM is compared with Probabilistic Neural Network (PNN) and it is observed that 70.19% and 78.73% of accuracy is attained by PNN and ELM respectively. In order to increase the accuracy of classification further, Extreme Deep Learning Machine (EDLM) and Complex Deep ELM (CEDLM) are used. Extension of an ELM from real domain to complex domain is known as Complex ELM. The performance of EDLM and CEDLM is measured by comparing it with ELM. After applying eco-friendly cleansing process, our proposed algorithms EDLM and CEDLM give the highest rate of performance measures of 85.87% and 92.05% when compared to ELM.

The test accuracy attains its maximum of 92.05% result which is better when compared to the results of the existing classification techniques like PNN, SVM and EDLM for collected copper plate Tamil character written images. In such cases, the CEDLM can improve the presentation of the system structure. Therefore, this proposal has increased the value for the field of Tamil Character Recognition precisely in the field of copper plate recognition.

This is extremely useful for researchers who are engaged in recognizing the metallic inscriptions worldwide as the same kind of metals can be found in most of the scripts used in the world.

9.2 FUTURE SCOPE OF RESEARCH

There is always a way better than the one that has been followed. Every versatile solution will have adequate flexibility for further extension. In any case, there are difficulties related with transcribed Tamil character acknowledgment, which has huge degree of scope for future research.

The composite characters have the essential structure looking like the consonants with minor alteration on the fundamental structure or have a supporting character which misleads to other characters.

Segmentation of content from non-content foundation is unexplored (for all the dialects) and has an incredible research potential.

Furthermore, many productive designs could be created for execution of Tamil character recognition. Using the same technique, recovery of characters is possible globally over any non-headline-based scripts.

In many places, it is prohibited even to take photocopy of the copper plates without incorporating eco-friendly cleaning processes. So, the data sample size was reduced to what was available. This research is a promising step towards bringing back vital information from even partially deteriorated copper plates which is otherwise left unattended. If this research is properly extended, which can be of use to research departments of archaeology and epigraphy, many precious copper plates information can be extracted and preserved for our future references.

DECLARATION

I, hereby declare that the thesis entitled **"SIMULATION TECHNIQUES TOWARDS RECOGNIZING TAMIL CHARACTERS INSCRIBED IN ANTIQUATED COPPER-PLATES: APPLICATION OF COMPLEX DEEP-LEARNING MACHINE ON ECO-FRIENDLY CLEANSED PLATES "** submitted to Bharath Institute of Higher Education and Research, Chennai for the degree of Doctor of Philosophy (Ph.D.) in Computer Science and Engineering, is the record of original research work carried out by me under the guidance of **Dr. M. PONNAVAIKKO** and has not formed the basis for the award of any degree, diploma, associate ship, fellowship or titles in this or any other University or any other similar Institution of higher learning.

Place: Chennai

Date: **(R.INDRA GANDHI)**

BONAFIDE CERTIFICATE

This is to certify that thesis entitled **"SIMULATION TECHNIQUES TOWARDS RECOGNIZING TAMIL CHARACTERS INSCRIBED IN ANTIQUATED COPPER-PLATES: APPLICATION OF COMPLEX DEEP-LEARNING MACHINE ON ECO-FRIENDLY CLEANSED PLATES"** submitted by **R.INDRA GANDHI** (**Reg. No: D15CS512**), for the degree of Doctor of Philosophy, Computer Science and Engineering , Bharath Institute of Higher Education and Research, Chennai–600 073, is based on the results of studies carried out by her under my supervision. This dissertation or any part of the work has not been submitted elsewhere for any other degree.

Place: Chennai **Dr. M. PONNAVAIKKO**

Date: **(Research Supervisor)**

CHAPTER 1

1. INTRODUCTION

1.1 OVERVIEW

India is known for its rich cultural heritage, from much before the prehistoric era. The lifestyle of Indians, their attitude, behavior, faith in poetic justice & religion, hospitality and literature were known from several sources such as inscriptions on temple rocks, walls, pillars, engravings in caves, copper plates, paintings, literature written on palm leaves and reports of foreign travelers etc. Epigraphically in-scripted monuments play a major role in knowing the past civilizations. Many civilizations were recognized only by the record of information that they left behind, with the help of their best linguistic potential. Among those in-scripted evidences, metallic monuments hold vast information. In particular, copper plate evidences play a prominent role in acting as the career of culture and civilization to the upcoming generations. A large portion of the recognized authentic copper monuments are influenced by climatic conditions particularly with long haul sedimentation of particles over earth, organic tainting, and so on. The negative impact is seen on majority of the parts present in open air climatic conditions, according to Knotkova (1999). Likewise, the impact on important monuments which are converted as tourist attractions over time and the social eradications like encroachments, stay homes built-up, shop expansions, prime spots enhancements, etc., play progressive impact on them. Old structures are the substance of urban networks that reflect the movements happened in a city over time, such as conflicts, wars etc., which are noted in the writings of Gonzalez (2007) and Saket Bhardwaj (2012). These even reflected the economic status of the city after sometime.

Securing old information can be considered as a way of reusing them, which reduces waste, saves imperativeness spent on amassing materials, gadgets etc., according to Huttenlocher (1993). With respect to culture, old copper plates help us understand history and enhance the respect for the people who lived centuries ago, practicing different traditions. Restoring copper plates require data and capacities more than those required to create fundamental measures and that is one of the inspiring powers for holding the identity of the past Architecture and Arts globally. As the historical copper monuments contain vital data, they deserve a special focus to be revived and preserved for our future followers.

1.2 BACKGROUND INFORMATION

India is a land of sound civilization and literature in which we come across an immense expansion of information when compared to the global content. In fact, Tamil is one of the old dialects of the world as its widely popular works of literature dates back over many centuries. Tamil, the local language of a southern state in India has few million followers around the world and it is an official language in nations like Sri Lanka, Malaysia, Singapore and Canada. There are innumerable evidences collected from these locations on the history of Tamil language and its heritage.

The origin of Tamil language dates back to many centuries and its inceptions are still not known. Yet, it has been created and extended in India as a language with rich writings. Tamil language has the biggest number of engravings in South Asia. The existence and the use of Tamil language can be traced back beyond pre-historic era and the language is

known for its aesthetics worldwide. It is the official dialect in the Indian province of Tamil Nadu and the association province of Pondicherry. All available ancient documents of Tamil language are accessible in any one of these forms which are stone inscriptions, metal carvings, palm leaf inscriptions and paper manuscripts. After careful analysis of historical metallic monument samples, particularly copper and bronze plates, it is acknowledged that all metal samples are being deteriorated because of corrosion. Copper Plate Character affirmation redesigns the treatment of copper plate pictures by allowing one to normally see and remove content substance from different data fields. In this aspect, copper plates which contain Tamil character deserve a special attention because of its age and sizeable contribution to the language and its literature.

"Advantage of any new system is the difference it can bring in the desired outcome. Even a marginal difference brought in brings a significant benefit towards the final outcome".

1.3 CHALLENGES IN HANDWRITTEN CHARACTER CLASSIFICATION

The practical benefits achieved from character recognitions, as well as the interests created when dealing with OCR problems directs one's interest towards further research in this area with measurable advantages in this field. The arena of Document Analysis and Recognition is widely used for character study in many areas. Character recognition is an important aspect when it comes to dealing with Document Analysis and Recognition. Character recognition is targeted on printed and handwritten documents. Off-line and on-line character recognition are the two important branches of handwritten character recognition. It is normally acknowledged that the on-line mode of identifying manuscript delivers better result than that of off- line manuscript identification mode.

In the present world, there are much OCR software available which can recognize a wide variety of characters, however, it is still difficult to recognize handwritten documents and fonts or scripts that mimic the handwritten works. All over the world, many researchers are trying out different approaches to achieve improvements in the techniques used on handwriting recognition. Improvements are being made to recognize characters based on the context of the word or sentence in which they appear. The accuracy is decided by various factors like document compositions, printing, copying and digitizing. Even a bit of digitization error that is negligible to human eye, may cause enormous decline in the accuracy of an OCR system.

This research is concentrated on Copper plate recognition. There are thousands of historical copper monuments that are slowly getting deteriorated because of poor maintenance and environmental factors. In the existing methodologies, chemicals which are used extensively to eradicate corrosion from copper objects are toxic and hence hazardous to the environment. Given below are few challenges being faced while adapting the existing methodologies and performing the task of character recognition on manually written documents,

- Complexity of characters partition from foundation
- Nonstandard types of images
- Nonlinear character segments
- Different characters have distinctive inclinations
- Neighboring images can be overlapping each other

- Some images might not be uniform
- Segmentation difficulties faced in recognizing contiguous characters
- Different manners in which a content is composed or periodic variations in the hand writing of authors

Our research work is primarily focused on overcoming these challenges. Restoring historical contents from deteriorating monuments is the main motivational factor behind this research work.

Among the listed problem statements above, the main focus areas covered in this research include:

- Segmentation is used to segregate the individual characters from manually written content which is a significant test in this framework. When a handwritten character is transcribed, most of the contiguous characters will in general be overlapping each other. Using Segmentation, we extract the individual characters from the overlapping characters for further processing.
- Different inscriptions composed even by the same author in different periods might not have the same strokes or shapes like the printed material. To overcome this limitation, we trained iteratively and set the data to identify the baseline character.
- The highlights utilized for preparing the classifier assume a significant job in the grouping. Selecting the appropriate element vector with respect to the types and

strips segregated can altogether improve the execution of a character classification. We have implemented vertical and horizontal projection technique to arrive at the appropriate character grouping.

1.4 PROPOSED METHODOLOGY

Historical monuments are vital documents in understanding our past. These are deteriorated over a period of time in which we have already lost a larger sum of information globally. Hence there is a necessity to study metallic monuments deterioration and to devise a solution to secure distortion less document processing for the recovery of corroded copper plates. Taking the environmental impacts as an important factor to consider, it is also necessary to come up with durable solution for environment-friendly corrosion removal techniques. With this aim, this research is primarily driven towards:

- Non-toxic Corrosion removal to preserve copper monuments information from deterioration.
- Distortion-less copper plate digitization.
- Character recognition using extreme deep-learning method.

The proposed work is concentrated on 11th century copper plates which are engraved in Tamil. The advantage of this strategy is to recognize various distorted contents written in the same copper plate engravings. This experimental work targets to implement a better methodology in improving the precision of grouping and the time taken for preparing the highlights separated from eleventh century handwritten Tamil contents.

The process of Character Recognition of any script can be broadly broken down into five proposed strategies given below which have been incorporated in our research to obtain the desired result:

Task I: Data Acquisition

Task II: Pre-processing

Task III: Segmentation

Task IV: Feature Extraction

Task V: Classification

1.4.1 Data Acquisition

Digitization is a method by which the grey-scale images are changed to binary images. The first step towards success in any image analysis or improvement program is its ability to categorize the objects of interest from the rest. Digitization splits the foreground (text) and background information. Local or adaptive threshold approaches apply different intensity values to dissimilar regions of the image. These threshold values are determined by the neighborhood of the pixel to which the threshold is being applied. The resultant digital images are then used as inputs to the pre-processing stage of data acquisition.

The primary goal of this exploration work is to characterize Tamil contents composed in digitized form of copper plate engravings. Epigraphists gained the specialty of interpretation of Tamil engravings. Despite the fact that epigraphists can peruse these old engravings, the information has not yet reached the common man. The inefficiency of the

common man to peruse and comprehend the engravings is the primary driving factor which has led numerous old landmarks to be neglected.

Examining these engravings, it is observed that numerous engravings were composed during the Chola regime of eleventh century. Engravings that were written during Chola regime utilizes various contents which helps us to understand various Chola administration practices that were prevalent during those periods. Hence, this exploration work is mostly focused on 11th century contents as there were many stone engravings and Copper plate engravings used during this period. An aggregate of 43 reported pictures containing 11th century inscriptions have been considered for this study. From each report a maximum of 350 characters can be reproduced which implies that roughly 12,900 characters can be reproduced.

1.4.2 Pre-Processing

The primary step in recognition of characters is the pre-processing stage. The main objective in the pre-processing stage is to eliminate all variations that may affect the final identification of the characters. This involves multiple operations on the digitized images intended towards reduction of noise and increasing the efficiency of removing structural features. Most of the historic copper plates identified are noticed to have been affected by environmental impacts like long-term exposure of these objects to dirt, biological contamination, etc. The poor effect is visible specifically in objects exposed to outside atmospheric environments. The natural climatic changes like change in temperature, precipitation, relative humidity, snow, gaseous and constant pollutions imposed by

business happenings also are predominant motives for this deterioration. Despite the fact that in indoor atmospheres these effects are much lesser, depositories or storage destructive corrosion may have formed on the copper object that can cause much more levels of damage to the subject. This work introduces an eco-friendly *phytochemical* method to remove corroded products from copper objects using *Bryophyllum calcynium* as main course material for corrosion removal along with supplementary binding agents.

Subsequently, the picture comprises of different dim levels. Thus, the threshold turns into a significant part in character recognition. This helps in simplifying the process and reduce the memory space utilization. This target system is utilized to extract explicit data that is required and to perform fundamental tasks by subsequently diminishing the impact of variations in penmanship. As the subjects are checked physically, there might be a chance of actual data getting misinterpreted. This may cause trouble in further processing of the picture. Thus, inclined recognition and adjustments must be incorporated to get genuine recognizable proof. The fundamental point of pre-processing is to process the pictures which are in crude structure and to get pictures which are reasonably cleaner for further processing.

1.4.3 Segmentation

Segmentation is the main aspect and an important phase in character recognition process. It is the progression of segregating the complete document image into recognizable units for feature extraction and classification. Text content from the input document is separated into multi-lines on which the further segmentation process is followed. Character segmentation is an important feature of an OCR system. For a character

segmentation system, it is important to identify the starts and ends of the characters. Character segmentation method helps to find gaps between words and spacing between adjacent characters.

Grouping of 11th century handwritten antique Tamil characters is the primary goal of this research work. To achieve this, character must first be extracted from the source picture which is an entire archive containing numerous lines. Hence, Division is implemented in this exploration which is a process of segregating the archive picture into content lines, words and subsequently into characters. Existing strategies utilize associated parts and projection profile systems for division. Inconveniences in those methods are that they separate straightforward characters into their constituent glyphs and when the characters are covered or contacted, they cannot be portioned appropriately. To overcome this, in our proposed strategy, the space between the lines is utilized to isolate the lines.

Ordinarily, the separation between two lines is bigger than the separations between words and in this manner, lines can be fragmented by looking at this separation against a reasonable edge. So, as to decide an ideal edge, PSO strategy is utilized.

Utilizing this, the base character is isolated from the modifiers. Furthermore, by utilizing the projection profile, the base character is divided first and afterward, utilizing the closest neighborhood strategy, the vowel modifiers and consonant modifiers are added back to the base character.

The fundamental step in character recognitions is segmentation process, which helps to isolate characters. Most of the existing character recognition systems lack in accuracy

because of input documents containing distorted text or overlapping characters, which leads to in-correct segmentation. The simple procedure of using inter-character gap for segmentation is usually useful for good quality printed documents, but this method fails to give agreeable outcomes if the input text contains touching characters. One of the main reasons for failure in character recognition of an OCR system is its deficiency in overcoming errors faced in character segmentation. As such, for an OCR system, segmentation process involves following steps:

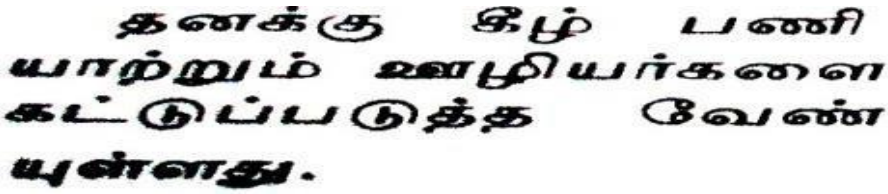
1. Segmentation of text regions from input document.
2. Segmentation of individual lines from text region.
3. Segmentation of individual words and zones from lines.
4. Segmentation of individual characters from word.

For our experimental purpose, we have selected only single column text without any graphic, tables, pictures, special symbol etc. As shown in Figure 1.1 below, we have followed the proposed segmentation process for distorted Tamil scripts in seven different steps.

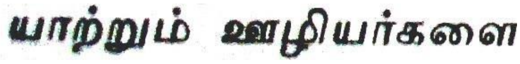

விடுமுறை நாட்களில் மன அழுத்தம் அதிகமாகிறது

(a) Sample Input document

(b) Extracted Text Region from Input

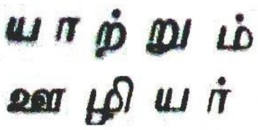


(c) Segmented Lines from text regions

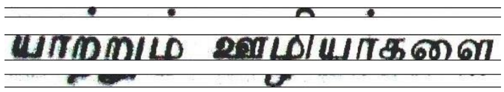


(d) Segmented Words from lines

(e) Segmented Zones from word



(f) Segmented Characters from word



(g) Segmented Matra from characters

Figure 1.1 (a) to (g) : Proposed Segmentation Processes for Distorted Tamil Scripts

Existing character recognition tools are adequate enough for recognizing characters without distortion. They mostly help to produce only alternate, wrong representation of the characters that are misinterpreted as the recognized characters, due to disturbances bundled with the source document. Proposed algorithm by Indra Gandhi et al (2010) helps to overcome this difficulty by applying the new technique, comprising various layers of segmentation based on the problems associated with the source document. In the following section, using horizontal and vertical projection technique we have formulated new algorithm for line, zone, and characters.

1.4.3.1. Line Segmentation

Line segmentation becomes a significant step in the segmentation process. Already, there are many techniques available for line segmentation but there are many limitations with these normally used techniques. In general, each text line is separated from the earlier and following lines by white space. That is, from the black and white image, horizontal black Pixel frequencies are calculated for every row. Then the lines are segmented by the rows having black pixel frequency of 0. As shown in Figure 1.2 the

segmentation strengthens the character recognition from distorted Tamil content as the horizontal projection of the document, which divides the whole document into various lines. In general, the line-height as well as symbol-gap are focused first. In the case of line segmentation, in text document average line, height is utilized to segregate pairs of text lines, which otherwise becomes difficult because of noise. In Figure 1.2 (a) and (b) encircled parts shows that there is some distortion in line segmentation for both hand written and printed document scripts. Tamil characters are usually a combination of 2 or 3 disconnected symbols hence, symbol is used to earmark individual components from the combination. To distinguish a word territory from a symbol boundary, symbol-gap statistics is used. To progress further, the individual symbols are separated by repetitive application of segmented word successive application of the morphological closing and connected component-based segmentation.

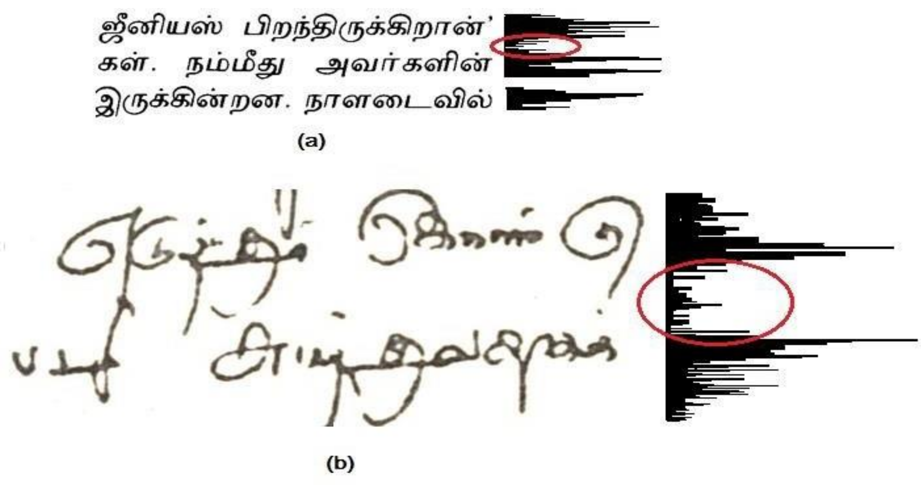


Figure 1.2: Horizontal Projections in Tamil Scripts

a. Same size printed script horizontal projection

b. Hand written script horizontal projection

The following are some general definitions and notations by Bansal (1999) which is used in our algorithms:

i). Horizontal Projection: For a given binary image of size L × M, where L is the height and M is the width of the image, then the horizontal projection HP(i), i = 1, 2, 3, …, L.

Where, HP (i) is the total number of black pixels in $i^{'th}$ horizontal row.

ii). Vertical Projection: For a given binary image of size L × M, where L is the height and M is the width of the image, then the vertical projection VP(j), j=1, 2, 3, ..., M.

Where VP (j) is the total number of black pixels in j^{th} vertical column

iii). Continuous Vertical Projection: For a given binary image of size L × M, where L is the height and M is the width of the image, the continuous vertical projection CVP (k), k = 1, 2, 3…M.

Where, CVP (k) counts the first run of consecutive black pixels in k^{th} vertical column.

Strip: A strip can be defined as a collection of consecutive runs of horizontal rows, each containing at least one pixel.

Horizontal projection on distorted documents of Tamil scripts divides the whole documents into strips which is listed in the following Table 1.1.

Table 1.1: Various Types of Strips Using Horizontal projection

Type No	Zonal Separation
1	Strips contains only upper (Matra alone)
2	Strips contains only middle (i.e., segmented from Upper and middle / middle and lower)
3	Strips contains only upper and middle (no lower)
4	Strips contains only upper and middle having lower zone (which is wrongly segmented as next lines upper)
5	Strip containing upper, middle and lower
6	Strips contains only lower (which is segmented from next lines upper)
7	Strip containing two or more overlapping lines
8	Strips contains only Middle and Lower (Upper Matra Segmented wrongly by previous line)

For example, Figure 1.3 is subjected over horizontal projection, which divides the whole documents into strips whose type identification is listed in Table 1.2.. From Figure 1.3 it is identified that there are 14 Lines of text. Uniform sized text of line 3 to 14 is taken as input data and it is mentioned as strip1, strip2 and it goes on up-to strip14. First line of text is eliminated from the test since it has different sizes when compared to other strips.

From the above Figure 1.3, it is clearly understood that there is no need to perform segmentation for the strips numbered 3, 12 and 13. Strips 1, 4 and 6 contain components of Matra part and it requires decision of the strip order which is needed to make complete line. Strip numbers 2 and 10 indicate the middle part and here also decision is necessary to match with upper or lower part of the line.

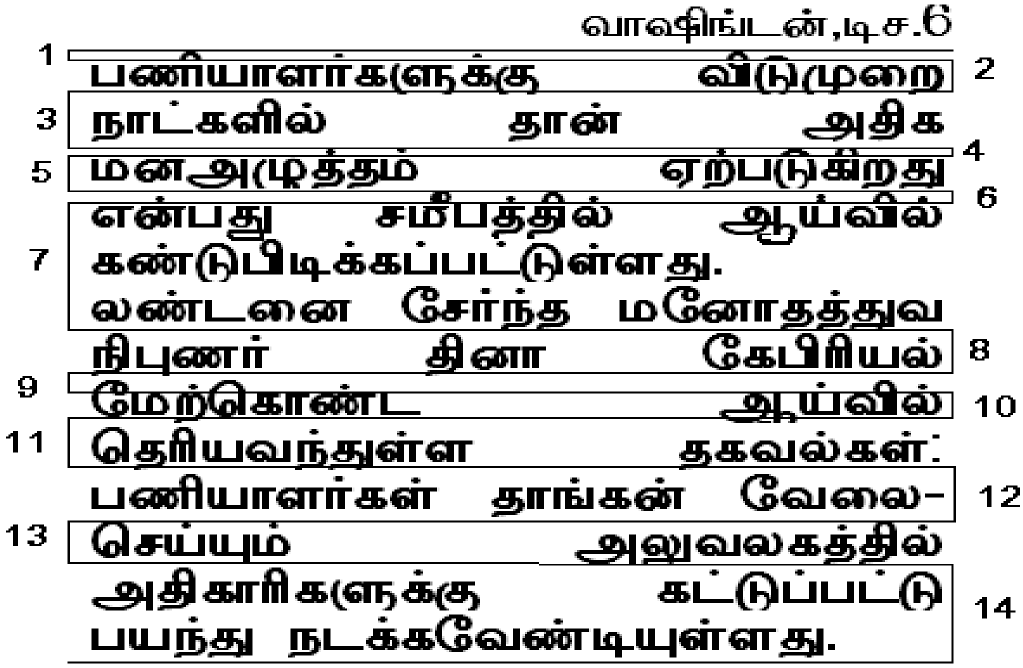


Figure 1.3: Various Types of Strips in Tamil Scripts

Strip number 5 shows incomplete line because of wrong segmentation by previous line. Similarly, strip numbers 8, 9 and 11 also have experienced wrong segmentation by the lines that follow. Strips 7 and 14 contain overlapping lines, which requires more attention and requires proper segmentation. As a result of this, it is necessary to find accurate boundaries (base line) for each strip.

There is a standard method suggested by Dholakia et al., (2005) for printed Gujarati characters. In this, the author has stated about the zone identification technique. This will not work successfully in all the cases of non-headline based Dravidian family scripts like Tamil, Malayalam, Telugu, Kannada etc., as these have different characteristics mainly structural difference when compared to Gujarati characters.

An extended algorithm for Tamil script is introduced here for identification of mean-line which is the base for further processing. This algorithm can be extended further to work with other Dravidian languages, based on their structural features.

Table 1.2: **Various Types of Strips Using Zonal Separation Techniques**

Type No	Strip No	Zonal Separation
1	1,4,6	Only upper (Matra alone)
2	2,10	Only middle (i.e., segmented from Upper and middle / middle and lower)
3	11	Upper and middle (no lower)
4	8	Upper and middle having lower zone (which is wrongly segmented as next lines upper)
5	3,12,13	Strip containing upper, middle and lower
6	9	Only lower (which is segmented from next lines upper)
7	7,14	Strip containing two or more overlapping lines
8	5	Middle and Lower (Upper Matra Segmented wrongly by previous line)

Considering the case of baseline, there are so many standard techniques given by Jindal et al (2006) and Garg et al (2010) that already exist. But we have used the following code for identification of baseline in non-headline-based scripts. For finding the mean-line and base-line, we have proposed two different algorithms which can be used as the base for the following line segmentation algorithm. In addition to this, it is used for zonal segmentation, which is carried out in the following section after word segmentation and before character segmentation.

1.4.3.2. Word and Zone Segmentation

The process of word segmentation follows the line separation task. Most of the word segmentation concerns regular core on the gaps among the characters to differentiate the words from each other. From this method of identifying words, it is found that the extraction of spaces among words is comparatively larger than the spaces between the characters. Isolation of word starts with recognizing the words from its textual lines. Features like strokes and concavity are used for decisive the segmentation points.

Figure 1.4: Vertical projection of a word

a) Hand written word – vertical projection

b) Printed character word – vertical projection

Inter word gap is utilized for effective word segmentation. We have also used the same technique as used by Varga (2006) for segmenting the words using vertical projection. Utilizing the same, a vertical projection profile is raised as shown in Figure 1.4. If in the

vertical projection profile when least k consecutive zeros are identified, that midpoint is considered as the margin of a word. Generally, the value of k is considered as half of the text line height. As shown in Figure 1.4, there is sufficient amount of gap in horizontal direction in the vertical projection of a line of distorted Tamil script for segmenting the words.

1.4.3.3. Segmentation Based on Zone Height (upper, middle, lower)

After segmenting the words, zone segmentation has been carried out. Zone segmentation is carried out based on the height of the horizontal project, which includes identification of middle, upper and lower zone boundaries. As already discussed in Fig 1.5, the region over the headline is called upper zone, the region from the headline to the baseline is called middle zone and region under the baseline is called lower zone. In general, text lines of any Tamil script are segmented into three different zones. For our reference, we name them as “Upper Zone, Middle Zone and Lower Zone”. The segments employed on an individual symbol are static and they continue persistently for the whole font. Almost 65% of the symbols in Tamil occupy the Middle Zone. Thus, the horizontal projection value of any row in the Middle Zone is large related to that of a row of the Upper Zone and Lower Zone.

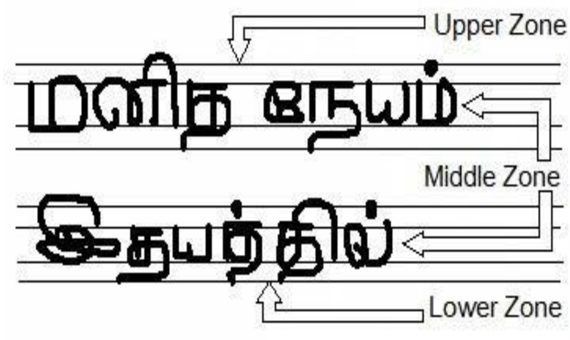


Figure 1.5: Contains all Three Segmented Zones of one word

1.4.3.4. Segmentation Based on Matra /Extensions

There is no literature survey available regarding the segmentation of Matra or Extensions for non-headline-based scripts. Some of the works done by Chaudhuri and Pal (1998), Lehal (2001) were based on headline-based scripts. Whereas Bansal (1999) shows his review in which the author has removed headline for preliminary segmentation. Special care is needed to segment distorted characters. We have developed some new matra segmentation algorithms based on the research of Indra Gandhi et al (2009) for normal / distorted characters of Tamil script in various zones. In this thesis, as shown in Figure 1.6 the matra has been classified into three different categories.

a) Matra / Extension in Upper Zone

b) Matra / Extension in Lower Zone

c) Matra / Extension in both Upper Zone and Lower Zone

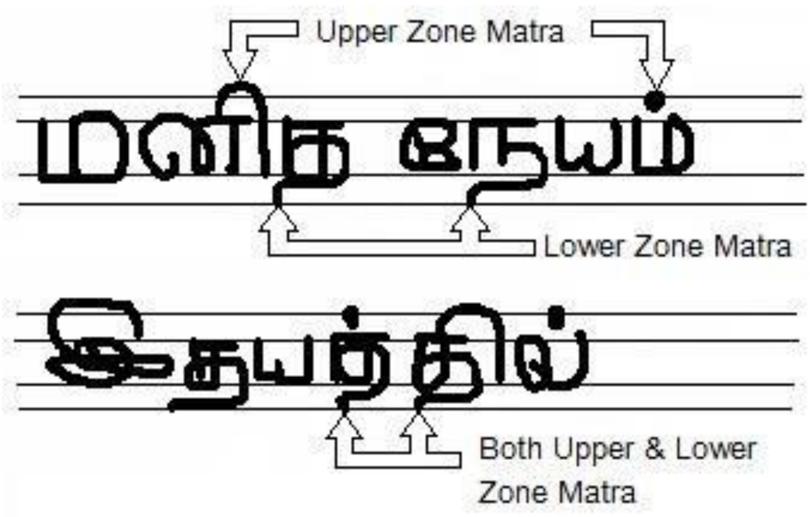


Figure 1.6: Different Zone Matra

1.4.3.4.1 Matra / Extension in Upper Zone

Based on structural features, the set of symbols that is present in upper zone is divided into Group 2, 3 and 4. In other words, the two disconnected components are the combination of Upper Zone and Middle Zone components. The structural feature classifications of upper zone Matra /extension are given in Table1.3.

Table 1. 3: Matra / Extension in Upper Zone

Group	Structural Features	Upper Zone Matra / Extension Name
2	Dotted	Pully “ • ”
3	Convex Shape	Kurill “ ”
4	Arc with small circle	Nedeill “ “

1.4.3.4.2 Matra / Extension in Lower Zone

Based on structural features, the set of symbols, that is present, is clustered into five groups namely Group 5, 6, 7, 8 and 9. In other words, the two disconnected components are the combination of Middle Zone and Lower Zone components. The structural feature classifications of lower zone Matra/extension are given in Table 1.4.

Table 1.4 Matra / Extension in Lower Zone

Group	Structural Features	Lower Zone Matra / Extension Name
5	Line	Kuril “ ”
6	Slanting Line	Kuril “ ” “ ”
7	Concave Shape	Kuril “ ” “ ”
8	Convex Shape	Kuril “ ”
9	Arc with small circle	Nedeil “ ”,

1.4.3.4.3 Matra / Extension in both Upper Zone and Lower Zone

Based on structural features, the set of symbols that is presents in both Upper Zone and Lower Zone are clustered into Group–9. The multi-component characters are the combination of upper, middle and lower Zone components. In other words, two components of upper and lower zone are joined with the middle zone components.

1.4.4 FEATURE EXTRACTION

Highlights must be extricated from the subjective old Tamil contents so as to isolate them into various classes. The Characteristic extractor generates some hard functions that allow the type to distinguish between styles of diverse characters. The function extraction selects a set of not unusual characteristic on the way to assist to uniquely perceive the character person. This selection set of features is the heart of sample popularity gadget layout. Various types of capabilities that have been used are edges, closed loops, strokes, and so on.

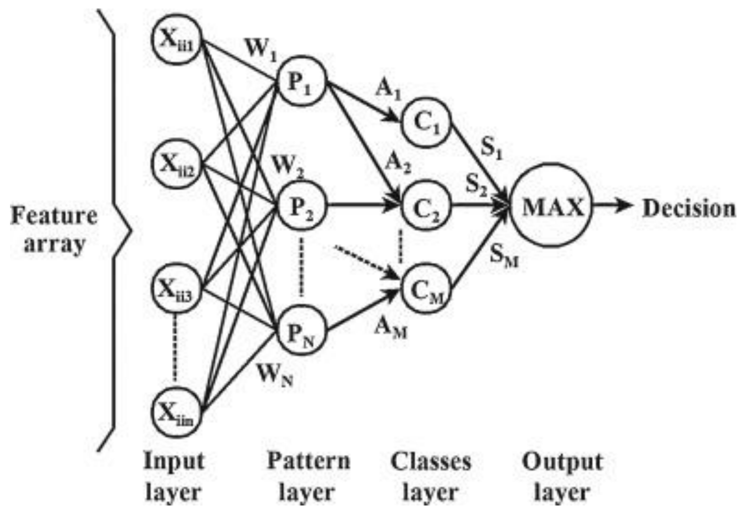


Figure 1.7: PNN Architecture

This exploration work mostly focuses on actual data and territorial highlights on the grounds that measurable highlights are invariant to character deformation and composing style to certain extent. Zernike minutes, Hu second invariant, relative second and local component are highlights which are generally utilized. These highlights are separated and framed into various component vectors. The best element vector is chosen by utilizing Probabilistic Neural Networks (PNN). From the exploratory outcomes, it is observed that including vectors containing Zernike minutes, Regional Features and mix of Zernike minutes and local highlights give better exactness in contrast to different vectors. Consequently, these 3 component vectors are mulled over for additional grouping reason. The PNN architecture is basically a back propagation network with an activation function derived from statistical data. The pattern and classes layers require supervised knowledge to connect each pattern layer node to the corresponding class layer node.

1.4.5 CLASSIFICATION

Classification is performed with respect to the class association of a pattern. That is, the classification task is to design a decision rule that is easy to compute based on the feature vector and the decision rule is implemented based on syntactical and statistical techniques. The classification enables reducing the possibility of mapping unknown character set against a subcategory of the total character set. To achieve such classification-based matching a selected domain is categorized into clusters and in such clustering the groups are not predefined as in classification. Hence, relevant algorithm will group similar items together.

Pertinent studies in this research include 11th century handwritten Tamil contents that are systematically gathered into four classes namely, Vowels, Consonants, Composite Characters and Special Characters (For example, Ayudha Elluthu in Tamil). These four classes are considered in arrangement procedure. Inasmuch as, customary calculations are slower than as needed, the slope-based learning and its parameter calculations are done iteratively.

Hence, an Extreme Learning Machine (ELM) is proposed. The proposed ELM is then contrasted against Probabilistic Neural Network (PNN). Pertinent studies reveal that, about 70.19% and 78.73% of precision is accomplished in using PNN and ELM respectively.

To improve the order of precision further, a more Complex version of ELM is suggested. The proposed Complex ELM is an extension of ELM into a complex space and the performance of the Complex ELM is decided by contrasting it with conventional ELM. The Complex ELM (CELM) enable highly elevated order of exactness (of about 80.30%) in contrast to traditional ELM.

To expand the exactness and diminish the number, shrouded neurons are utilized. To diminish the time taken for preparing, an advanced Complex ELM has been proposed in this exploration work. In complex ELM, the info loads and shrouded inclinations are haphazardly produced. This may prompt arrangement of non-ideal information loads and shrouded predispositions. So as to compute an ideal information weight and shrouded inclinations, Differential Evolution (DE) calculation is utilized with Complex

ELM. DECELM, when prepared utilizing highlight vector acquired by joining the Zernike minutes and territorial highlights, gives the most elevated exactness pace of 83.27% when contrasted with ELM and CELM. Moreover, the preparation time is diminished when contrasted with that of including CELM.

1.5 FRAMEWORK OF THE PROPOSED METHODOLOGY

CPCR (Copper Plate Character Recognition) is the acknowledgement of composed Tamil content characters from a copper plate. This includes photographed examination of the content character-by-character, investigation of the filtered picture and interpretation of the character picture into character codes, for example, ASCII, normally utilized in information formulating. A considerable lot of present CPCR frameworks are assembled in customary ways to deal with prepared pictures and works incredibly with printed messages. Utilizing them for handwritten content in pictures it can give surprising outcomes with poor acknowledgment quality.

In the underlying phase of pre-processing, binarization and slant discovery and amendment are the two pre-processing strategies used in this exploration work as shown in Figure 1.8. In binarization, Otsu, Kittler Met and Niblack calculations are utilized and their exhibits are assessed utilizing their measurements.

For example, PSNR, SNR, MSE, Precision, Recall and F-measure are used. From the assessment, it is observed that Otsu calculation gives better outcome when compared with rest.

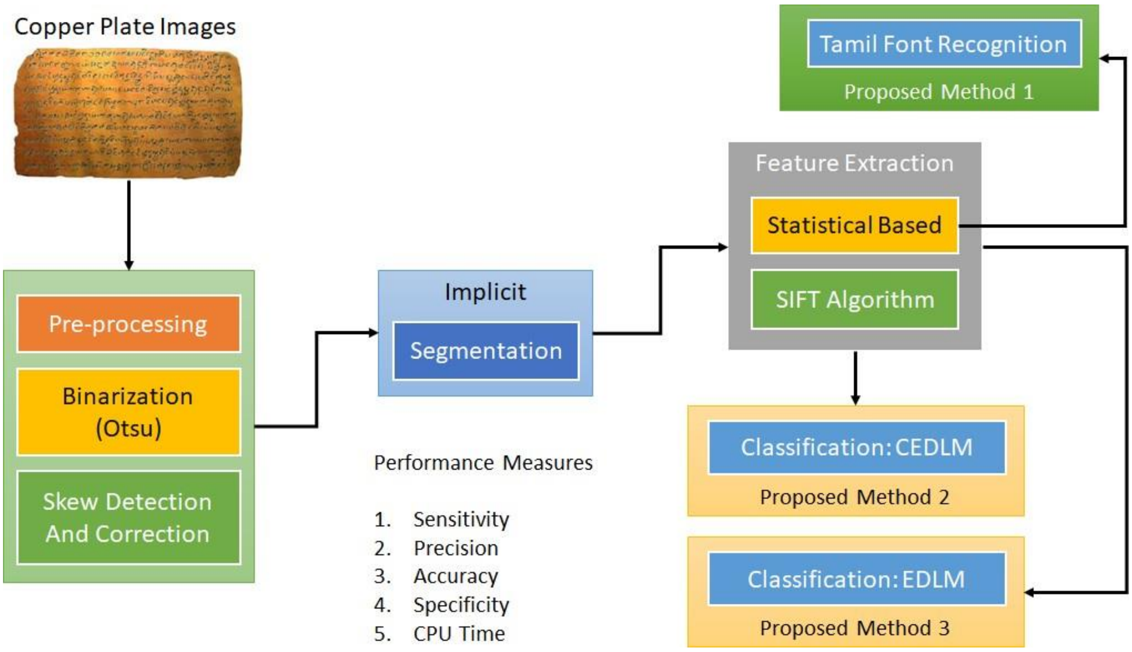


Figure 1.8: Framework of the Proposed Methodology

To show signs of improvement on binarized picture, Otsu calculation is altered by utilizing molecule swarm advancement strategy. The altered strategy on Otsu is contrasted with Niblack and Kittler Met calculations. The trial result shows that the proposed calculation gives better binarized picture when contrasted to different calculations.

The resultant work is characterized further utilizing Particle swarm, streamlining the target work which gets boosted to distinguish the best slant point. The results of changed projection profile is contrasted with Hough change, Fourier descriptor and projection profile techniques. For grouping of 11th century transcribed Tamil contents, Extreme Learning Machine (ELM) and Complex Extreme Learning Machine (CELM) are utilized. Further to build the exactness of the arrangement, CELM is upgraded by utilizing Differential Evolution (DE-CELM). Its exhibit is estimated by contrasting

CELM and ELM. The test results show that the proposed and improved complex machine learning arranges the eleventh century handwritten antiquated Tamil contents with noteworthy precision rate when contrasted with ELM and CELM.

1.6 ORGANISATION OF SYNOPSIS

The underlying objective of this research work is to classify 11th century handwritten Tamil scripts using computational intelligence technique. This report consists of nine chapters.

Chapter 1 presents the introduction about 11th century scripts and the need for using these scripts for classification along with the objectives of the proposed work. It also briefly explains the various tasks of the proposed methodology and its overall framework.

The literature analysis is a noteworthy and evaluative summary of specifically defined research topic gained from the available collected works. Chapter 2 describes the existing literatures available in the classification of handwritten Tamil scripts and also describes and analyses the issues related to present research works.

Chapter 3 describes the research methodology followed by discussions on pre-processing techniques carried out in the research work and explains in detail about the proposed algorithm.

Chapter 4 explains the historical Metallic Monuments Preservation Techniques against Deterioration and the tendency to introduce an eco-friendly phyto-chemical technique

for the removal of corrosion merchandise from copper objects victimization.

Statistical Feature Extraction Based Copper Plate Character Recognition is proposed to build a suitable mode to rescue valuable ancient inscriptions, in a convenient way is discussed in Chapter5.

Chapter 6 explores the best feature extraction methods and selecting a proper selection algorithm and classification techniques which leads to established recognition accuracy and low computational overhead. This chapter proposed, a new methodology of Extreme Deep Learning Machine (EDLM), algorithm for classification that has a short processing time is proposed.

Feature extraction plays an imperative role in the classification of copper plate Tamil font characters recognition techniques together with proposed Complex Extreme Deep Learning Machine (EDLM) feature vectors are discussed in chapter 7.

In chapter 8, the proposed EDLM and CDELM classifier results and comparison of results of the existing classifier with the proposed classifier are tabulated and discussed in detail.

Finally, chapter 9 presents a discourse on the overall conclusion and future works of the suggested research work. The works of earlier researchers are mentioned and used as suggestion to hold up the thoughts explained in this dissertation. All such proofs used are listed in the reference section of the dissertation.

1.7 SUMMARY

This chapter provides a brief overview to the research problem, that is, classification of 11th century handwritten Tamil scripts. The purposes formulated are also outlined. To achieve the objectives outlined in this chapter, a review of the previous research work is done and the related works are summarized in the next chapter which is the review of literature.

CHAPTER 2

2. LITERATURE SURVEY

"The heritage of the past is the seed that brings forth the harvest of the future"

2.1 INTRODUCTION

This chapter provides a comprehensive study of the research outcome in several related concepts and methods which are used in this thesis. The first part reviews on world dialects (Non-Indian Script), the second part gives an exploring view to the works on Indian dialects, Indian Script Recognition and then narrows it down to the works on character acknowledgment to local language (Tamil Character recognition). Finally, the review was carried out on the core area of existing technique towards evolutions and edge detection challenges in character recognition.

2.2 RESEARCH WORK ON NON-INDIAN SCRIPT RECOGNITION

The starting point of incredible arrangement of study work in mid-sixties depended on a methodology known as examination by synthesis strategy proposed by Cox et al (1974). The incredible significance of Eden's work was that he officially demonstrated that every single manually written character are shaped by a restricted number of schematic highlights, a point that was verifiably remembered for the past works. This

thought was later consumed in all procedures in syntactic (auxiliary) approaches of character acknowledgment. In sixties, Narasimhan (1964), Narasimhan (1966) recommended a naming pattern for Syntactic portrayal of pictures, and a sentence structure coordinated understanding of classes of pictures. In another work, Narasimhan (1969) proposed an acknowledgment strategy dependent on depiction and age. Utilizing natives and relations, he depicted a particular language for hand-printed FORTRAN character acknowledgment. Later Narasimhan et al (1971) set forward a sentence structure that helped in acknowledging combine, wherein they have recommended that the principles 34 at present being used must be refined and adjusted persistently based on the understandings and other information gained. Pavlidis et al (1975) and Ali et al (1977) recommended part and consolidation calculation for polygonal estimation of characters for numeral acknowledgment. A component age system for syntactic example acknowledgment by approximating character limit utilizing polygons and decaying based on concavity is recommended by Feng et al (1975).

Significant research work in character acknowledgment is currently focused on manually written Chinese characters, which is as yet viewed as an exceptionally difficult issue and viewed as extreme objective of acknowledgment investigate Leung et al (1998), Yamamoto et al (1984). Obviously, the early work was on hand printed characters. Casey et al (1996) at IBM introduced one of the main endeavors in Chinese character acknowledgment.

Agui et al (1979) proposed a depiction technique for hand-printed Chinese characters acknowledgment. Later in 1981, Fuji et al (1981) showed a model for hand-printed

Kanji character recognizer and the mental boundary was broken. From that point forward, a lot of work has been done in Japanese and different existing languages just as new strategies have been attempted.

Arakawa (1983) recommended an on-line manually written character acknowledgment framework for Japanese characters. Another unwinding technique dependent on highlights mirroring the basic data of Chinese characters is presented by Xie et al (1988).

An altered unwinding technique has been recommended by Leung et al (1998). Recommended acknowledgment by means of neural systems for accomplishing quick acknowledgment of hand-printed Chinese characters.

Zhao (1990) has presented the plan and acknowledgment of two-dimensional broadened property syntax strategies for the acknowledgment of hand printed Chinese characters. A technique dependent on the idea of Fuzzy set for transcribed Chinese characters is proposed by Cheng et al (1985). A tale stroke-based component to remove skeletons and basic 35 highlights of Handwritten Chinese character acknowledgment is proposed by Chiu et al (1999). Different procedures utilized for acknowledgment of Chinese typescripts can be found in the analysis work of Mori et al (1984), Tappert et al (1990) and Wakahara (1993).

Measurable and basic strategy-based acknowledgment of Cursive Arabic characters acknowledgment is accounted for in the works of El-Dabi et al (1990) and Almuallim et al (1987). A stroke-based disconnected and transcribed Korean characters work is

accounted for in Kim et al (1996). Acknowledgment of hand written characters has been concentrated well in the writing similar to Chinese, Arabic and a couple of contents of the countries that created them are concerned. Studies of related works are found in Srihari et al (2006), Amin (1997), Lorigo et al (2006), Stefan Jaeger et al (2002).

Acknowledgment of Chinese characters utilizing second descriptors is introduced in Simon Liao et al (1996). Neural system-based calculation for perceiving Chinese characters was given in Mingrui Wu et al (2000). In Feng Lin et al (2002) a quick and precise calculation to extricate the strokes from the diminished Chinese character pictures was depicted. An approach for disconnected written by hand Chinese character acknowledgment, dependent on converging from back-to-back the sections of versatile length, is introduced in LI Guo–hong et al (2004). Disconnected written by hand character acknowledgment framework for English language dependent on investigative methodology, neural arrangement of characters is introduced in Hariton Costin et al (1998).

In Hanmandlu et al (1999), a combination of ring-based and segment-based technique for the acknowledgment of written by hand English capital letters was proposed. A classifier for Arabic character pictures was structured utilizing a choice tree enlistment calculation and Multilayer Perceptron (MLP) organized in Mustafa Syiam et al (2006).

2.3 RESEARCH WORK ON INDIAN SCRIPTRE COGNITION

Sethi et al (1977) has introduced a Devanagari numeral acknowledgment wherein the nearness or nonattendance of 4 essential natives, for example, level, vertical, right

inclination, and left inclination are utilized for acknowledgment with the assistance of choice tree. Later, they endeavored to perceive imperative hand-printed Devanagari characters utilizing a similar technique.

Sinha et al (1985) completed Devanagari content acknowledgment utilizing syntactic technique with an installed picture language. It is a model-setting-based acknowledgment. Sinha (1987) later recommended role of setting in Devanagari content acknowledgment framework. Siromoney et al (1978) endeavored machine acknowledgment of Printed Tamil characters utilizing an encoded character string word reference. Later, they proposed an acknowledgment method for printed Brahmi. The plan depends on run-length technique. Comparable methodology is applied to requirement hand-printed Tamil characters by Chandrasekaran et al (1984) for multi-font Tamil and extraordinary arrangements of Printed Malayalam and Devanagari characters.

Chinnuswamy et al (1980) introduced a methodology for hand-printed Tamil character acknowledgment utilizing named diagrams to portray auxiliary synthesis of characters as far as line-can imagine natives. Acknowledgment is completed by connection coordinating of the named chart of the obscure character with that of the model. A two-phase acknowledgment framework for Telugu letters in order has been portrayed by Rajasekaran et al (1977).

The primary stage depicts the coordinated bend following technique with an information-based inquiry to perceive natives 37 and to remove essential character from the real character design. The subsequent stage portrays the coded essential

character and its acknowledgment through choice tree. An endeavor for Malayalam character acknowledgment is conveyed by Janarthanan (1993). In this approach, three level grouping plans is utilized for acknowledgment reason, as angle proportion, Fourier descriptor coefficients and number of natives. The last rectification is completed by semantic guidelines.

Nagabhushan et al (1999) have proposed a non-uniform measured Kannada characters acknowledgment utilizing a locale disintegration and ideal profundity choice tree technique. An endeavor for Bengali character acknowledgment was taken up by Ray et al (1984). They introduced a closest neighbor classifier utilizing highlights separated by utilizing a string availability standard. Misusing the closeness among the significant Indian contents, Dutta (1984) introduced a summed up formal methodology for age and examination of all Bengali and Hindi characters.

Marudarajan et al (1978) utilized versatile limit rationale for printed Hindi numeral acknowledgment. Sural et al (1999) have developed Bengali content acknowledgment utilizing fluffy component extraction dependent on Hough change of Multi Layered Perceptron. There were no adequate number of studies on Indian language character acknowledgment. A large portion of the bits of existing work are worried about Devanagari and Bangla content characters, the two most famous dialects in India. A few investigations are accounted on for the acknowledgment of different dialects like Telugu, Oriya, Kannada, Punjabi, Gujarati and similar languages. On Printed Devanagari content, OCR work began on mid-1970s.

In Veena Bansal et al (2001), a total technique for division of content imprinted in

Devanagari was 38 introduced and the creator utilizes the basic properties of the Devanagari content.

In Reena Bajaj et al (2002), multi-classifier design had been proposed for expanding dependability of the acknowledgment consequences of Devanagari numerals. Directed and solo learning frameworks are joined to perceive manually written Devanagari numeral acknowledgment Patil et al (2007). A framework towards the acknowledgment of disconnected hand written characters of Devanagari, the most mainstream content in India is proposed in Pal et al (2007). The highlights utilized for the acknowledgment reason for existing are for the most part dependent on directional data got from the bend digression of the slope. In spite of the fact that examination on Bangla character acknowledgment began in mid-1990s, no huge work was accounted till mid-1990s. Of late, a few bits of work on Bangla have been distributed.

Chaudhuri et al (2007) proposed the Support Vector Machine (SVM)-based way to deal with Bangla character. In Sameer Antani et al (1999) the creators portrayed characterization of a subset of printed Gujarati characters. For the grouping, least Euclidean separation and K-closest neighbor classifier were utilized with ordinary and invariant minutes. A Hamming separation classifier was likewise utilized. The acknowledgment pace of the revealed framework was exceptionally low.

In Arun Pujari et al (2002), a strategy was proposed to perceive a Telugu content which utilizes wavelet multi-goals examination for the reason removing highlights and acquainted memory model for the acknowledgment task. In Rahman et al (2002), a multistage approach is proposed to perceive transcribed Bengali characters.

Improvement of OCR framework for printed Oriya content was troublesome on the light of the fact that an enormous number of character shapes and numerous indistinguishable characters are available in the content. Besides, round state of the vast majority of the characters has additional issues in the acknowledgment procedure. Just a couple of bits of work have been accounted for, on the acknowledgment of Oriya characters.

A framework was produced for the essential characters of Oriya content. Different pre-processing activities were done on the report picture. Next, singular characters are perceived utilizing a mix of stroke and run number-based highlights, alongside highlights acquired from the idea of water flood from a supply. The proposed framework by Pal et al (2003) perceives individual printed Urdu characters utilizing a blend of topological, shape and water store idea-based highlights. In the ongoing paper, Manjunathetal (2008) a multilingual character acknowledgment framework for printed South Indian contents (Kannada, Telugu, Tamil and Malayalam) and English reports was proposed. The proposed multilingual character acknowledgment depended on Fourier change and head segment examination.

2.4 COPPER PLATE TAMIL CHARACTER RECOGNITION

Copper Plate Character affirmation redesigns the treatment of copper plate pictures by allowing one to normally see and remove content substance from different data fields. Tamil character portion from Copper plate is a noteworthy assignment for acknowledgment System. Segmentation is the method for parceling the picture into content lines, words and subsequently into characters which are particularly useful for

arrangement. Knotkova et al (1994-1999) elaborated separation of character from Copper plate unique duplicate is testing, while the characters structure and substance differentiate inside and out. The precision of the OCR system depends upon the segmentation. If the characters are segmented precisely, the acknowledgment structure gives best results. Locales or segments are parceled from an image in division stage. Overwhelmingly, the division endeavors to isolate principal part of the substances which are undeniably characters. This is alluring by considering the way that the classifier sees these characters just. Segmentation stage is moreover essential in adding to this error in view of reaching characters, which the classifier cannot successfully see. Undoubtedly, even in incredible quality reports, some near to content style get in touch with one another due to the less ideal checking objectives expressed by Pago (1984).

2.5 WHY 11TH CENTURY SCRIPTS

Tamil contents are fundamentally advanced from the Grantham content around the seventh century Common Era (CE). Engravings, in this content are found in the northern segment of Tamil Nadu in the start of the eleventh century. Just during the eleventh century, engravings in Tamil contents came to use in the outrageous southern segment of Tamil Nadu (http://www.tnarch.gov.in/epi/ins2.htm). After this, palm leaves and stone engravings are turned into the prime media of composing. In this manner, there could be numerous writings and a therapeutic note that has been composed on palm leaves and engravings. In Tamil culture, the time of the Cholas (850 AD–1200 AD) was the brilliant period, which is set apart by the significance of writing. The Tamil nation arrived at new statures of predominance in workmanship, religion and writing under the Chola administration.

Tamil creators were motivated by the historical backdrop of Cholas as it offered an approach to deliver abstract and aesthetic manifestations during the most recent and a very long while study. Their support of Tamil writing and their energy in the structure of sanctuaries have brought about some incredible works of Tamil writing and engineering. Chola's engravings refer to numerous works (http://schools-wikipedia.org/wp/c/Chola_Dynasty.htm), dominant part of which has been lost, henceforth, so as to spare the forgotten about the engravings of the incredible work done during Chola's period, which have been composed utilizing eleventh century contents, that impact on the digitization of eleventh century written by hand contents are finished.

Individuals from different south Indian imperial traditions utilized Tamil copper-plate engravings to record their awards. The awards go in date from the tenth century CE to the mid nineteenth century CE. A significant piece of the commitment towards the old Tamil contents was given by the Chola lords. These plates are epigraphically significant as they provide us with an understanding about the social states of medieval South India and help fill ordered holes to interface the historical backdrop of the decisions of traditions. In the Chola line, numerous engravings were composed utilizing antiquated Tamil contents which contain notes about writing, medication, crystal gazing, and homeopathy, etc. Getting and keeping up these data are troublesome. On the off chance that these engravings are digitized, rich Tamil substance can be obtained by individuals of differing classes easily and comfortably. This lays the principal purpose for picking eleventh century Tamil contents for characterization. Table 2.1 and Table 2.2 shows an examination between 11^{th} century antiquated Tamil contents and current Tamil contents.

Table 2.1 : Uyireluttu of 11th century Tamil scripts

Current Tamil Scripts	அ	ஆ	இ	ஈ	உ	ஊ
11th Century Tamil Scripts						
Current Tamil Scripts	எ	ஏ	ஐ	ஒ	ஓ	ஔ
11th Century Tamil Scripts						

Table 2.2: Meyyeluttu of 11th century Tamil scripts

Current Tamil Scripts	க	ங	ச	ஞ	ட	ண
11th Century Tamil Scripts						
Current Tamil Scripts	த	ந	ப	ம	ய	ர
11th Century Tamil Scripts						
Current Tamil Scripts	ல	வ	ழ	ள	ற	ன
11th Century Tamil Scripts						

2.6 RECENT RESEARCH WORK IN TAMIL CHARACTER RECOGNITION

Tamil is one of the widely spoken Dravidian languages, spoken predominantly by several million speakers. It has gained official status in India, Sri Lanka Malaysia and Singapore. Minorities in Mauritius, Vietnam and Reunion also speak Tamil. Very recently, Indian Government has recognized it as a classical language. Despite a certain degree of influence by Sanskrit, Tamil stands unique away from the descendants of Sanskrit such as Hindi, Bengali and Gujarati. In the immediate past, international Tamil community has started extensive use of Tamil in computers.

Tamil is a standard language which is broadly spoken in most bit of the south India. There are 12 vowels, 18 consonants and one excellent character present in striking Tamil Script. Each vowel joined by unadulterated consonant to make an estimation of 216 consonant-vowel (CV) mixes.

These connote an aggregate of 247 Tamil characters. Tamil Language alphabetic system has been taken from the antiquated Brahmi content which fills in as a base for most of the Indian lingos. The vowels and consonants of Tamil letters set all together are given in the Table 2.3 and scientific categorization of arrangement methods are listed:

Table 2.3 Modern Tamil Character Set

Vowels	அ, ஆ, இ, ஈ, உ , ஊ ,எ , ஏ , ஐ , ஒ , ஓ , ஔ
Constants	க், ங் , ச் , ஞ் , ட் , ண் ,த் , ந் , ப் ,ம் , ய் ,ர் ,ல் , வ் ,ழ் , ள் , ற் , ன்
Grantha	க்ஷ, ஷ், ஸ்ரீ
Aytam	ஃ

Siromoney et al (1978) portrayed a technique for acknowledgment of machine printed Tamil characters utilizing an encoded character string word reference. The plan utilizes string highlights removed by line and segment savvy checking of character grid. The highlights in each line (segment) are encoded reasonably relying on the intricacy of the content to be perceived.

Chinnuswamy et al (1980) have proposed a methodology for hand-printed Tamil character acknowledgment. Here, the characters are thought to be made out of line-like components, called natives, fulfilling certain social limitations. Marked charts are utilized to portray the auxiliary arrangement of characters as far as the natives and the social requirements are fulfilled by them. The acknowledgment method comprises of changing over the info picture into a marked diagram speaking to the information character and registering relationship coefficients with the named charts put away for a lot of essential images.

Suresh et al (1999) endeavors to utilize the fluffy idea on manually written Tamil characters to order them as one among the model characters utilizing a component called good ways from the edge and a reasonable 46 participation work. The model characters

are arranged into two classes: one was considered as line characters/designs and the other was circular segment designs. The obscure information character was ordered into one of these two classes first and afterward perceived to be one of the characters in that class.

Aparna et al (2004) proposed a technique to build a written by hand Tamil character by executing a succession of strokes. A structure or shape-based portrayal of a stroke was utilized in which a stroke was spoken to as a string of shape highlights. Utilizing this string portrayal, an obscure stroke was recognized by contrasting it and a database of strokes utilizing an adaptable string coordinating methodology. A full character was perceived by distinguishing all the part strokes. Correlation of flexible coordinating plans for essayist subject to line 45 penmanship acknowledgment of detached Tamil characters was given by Niranjan Joshi et al (2004). Dynamic Time Warping (DTW) for ordering written by hand Tamil characters was given in Ralph Niels et al (2005).

Indra Gandhi et al proposed [2009] another technique of using Kohonen SOM (Self Organizing Map) to see the online Tamil character. The vectors of the twofold picture are made. Exactly when the division of the character is done, the photos are scaled to remarkable height and weight. Some unwanted fragments are consolidated, yet it will in general be cleared by sobel edge recognizable proof. The center channel is used to extend the efficiency. The SOM is not material to the cursive characters which are used at this moment.

Jagadeesh Kannan et al (2008) used Octal Graph procedure for the affirmation of the Tamil Handwritten characters. A complex approach of script distinction in each row or column which are encoded suitably for the script recognition was made use of, for constrained hand-printed Tamil character recognition, according to Chandrasekaran *et al* (1983). An approach of it deals with hand-printed Tamil character recognition employing labelled graphs to describe structural composition of characters in terms of line-like primitives. A brief explanation proposed by Dhamayanthi and Thangavel (2000) about the shape of the Tamil characters and Hewavitharana and Fernando (2002) extended a system to recognize handwritten Tamil characters using a two-stage classification approach, for a subset of the Tamil alphabet, which is a hybrid of structural and statistical technique. The idea of Seethalakshmi *et al.* (2005) is that which raises a dearth of OCR of Tamil character that should be digitized for sharing the data through Internet, which will enhance the process of convection of ancient and old documents into the latest.

Bharath and Sriganesh (2007) proposal of data-driven online handwritten word recognition system for Tamil complemented HMM-based word modelling. Yet another attempt was made for printed Tamil characters by Loganath and Shivsubramani *et al* (2007), using Multi-class Hierarchical SVM that constructs the hyper-plane to separate each class of data from other. Another work by Jagadeesh and Prabhakar (2008) was noted to increase the performance of Tamil OCR using the fusion algorithm. Venkatesh (2009) incorporated supervised learning algorithm using Support Vector Machine (SVM) for the recognition of handwritten Tamil characters. One more contribution by Venkatesh and Sureshkumar (2010) using back propagation network provide good recognition accuracy of handwritten Tamil characters. New encouraging methodologies are specified from the view purpose of general example acknowledgment technique.

Swethalakshmi et al (2006) a framework for acknowledgment of on-line manually written characters had been introduced for Devanagari and Telugu contents. A manually written character was spoken to as an arrangement of strokes and its highlights are removed and grouped. Bolster vector machines have been utilized for building the stroke acknowledgment motor. A tale cross breed written by hand recognizer was proposed for Tamil words were introduced in Srinivasagan et al (2006).

Hewavitharana et al (2002) portrayed a framework to perceive manually written Tamil characters utilizing a two-phase arrangement approach, for a subset of the Tamil letter set. In the main stage, an obscure character was pre-classified into one of the three gatherings: center, climbing and diving characters. At that point, in the subsequent stage, individuals from the pre-arranged gathering are additionally broken down utilizing a measurable classifier for definite acknowledgment.

A similar work on handwritten Tamil Characters by Suresh *et al* (2005) utilized the fuzzy concept. Author proposed a framework to perceive printed characters, numerals and manually written Tamil characters utilizing Fuzzy methodology.

Patil et al (2007) proposed a way to deal with utilizing the fluffy idea to perceive written by hand Tamil characters and numerals. The written by hand characters (numerals) are pre-processed and portioned into natives. These natives are estimated and marked utilizing fluffy rationale. Strings of a character are shaped from these marked natives. To perceive the written by hand characters, traditional string coordinating was performed. In any case, the issue in this string coordinating had been abstained from utilizing the enrolment estimation of the string.

Bhattacharya et al (2007) proposed a two-phase approach. In the primary stage, a solo bunching technique was applied to make fewer gatherings of written by hand Tamil character classes. In the subsequent stage, a managed characterization procedure was considered in every one of these littler gatherings for definite acknowledgment. The highlights considered in the two phases are unique.

A particularly made and physically composed character Recognition structure is not until now available for Tamil language. The guideline purposes behind this are:

(i) Tamil Language has an amazingly enormous character set.

(ii) Letter structure are flighty.

(iii) As a result of complex letter structure, making styles out of people change by and large.

(iv) There is no Tamil character database that exists for testing purposes in the open zone. Most, by far of the perceived unquestionable copper plate contents, are mainly affected by climatic condition; especially with long stretch presentation of articles over earth, natural contamination, etc.

Sutha and Ramaraj (2007) proposed a framework to perceive manually written Tamil characters utilizing Neural Network. Fourier Descriptor was utilized as the element to perceive the characters. The framework was prepared utilizing a few unique types of penmanship given by both male and female members of various age gatherings. The above writing study demonstrates that the current research chips away at transcribed Tamil character acknowledgment depended distinctly on Fuzzy methodology, neural system and factual methodology. Table 2.4 listed various taxonomy of classification techniques, as of late Hidden Markov Models has gotten consideration for character acknowledgment. To the best of the analyst's information, there is constrained work dependent on HMM to perceive written by hand characters. Just now, analysts are

structuring frameworks and growing new procedures for cursive written by hand characters utilizing Hidden Markov Models, Neural Networks, Fuzzy strategies and Neuro-Fuzzy methods and blend of these and so on. At present, Shin et al (1999) view on Genetic Algorithm (GA) is seen as the most amazing and reasonable improvement technique for reviewing an enormous clarification, and is used to find the most streamlined and ideal answer for a given issue.

Zhu and Garcia-Frias (2004) proposed two novel generative strategies which utilize stochastic setting free language structures and HMM's individually to display the start to finish mistake profile of radio channels. In any case, in their methodology, the structure has not been found out naturally inside a solitary probabilistic system. Another promising methodology that can contribute in building basic shrouded Markov models is that of Geman's work in vision. He acquainted compositionality as a capacity with developed progressive portrayals of scenes, whereby constituents are seen in an interminable assortment of social pieces.

Table 2.4. Taxonomy of Classification Techniques

Algorithm	*Advantages*	*Disadvantages*
Support Vector Machines (SVM)	• Flexible to handle classification and regression tasks of varied complexities. • Automatically select their model size.	• SVM is not very scalable in dealing with large data. • Long training and testing time.
Artificial Neural Networks (ANNs)	• Adaptive, robust, non- linear. • Generalization. • Can learn multiple outputs at the same time.	• The training time is relatively long. • Susceptible to local minimum traps. • Cannot be retrained.
Fuzzy Cluster Algorithm	• High Accuracy. • Flexibility. • Interpretability.	• Small training sample • Number of clusters must be specified beforehand.
Evolutionary Algorithm (EA)	• Very easy to understand and does not demand the knowledge of mathematics. • Easily transferred to existing simulation. • Variation prevent trapping in one part of the solution space.	• Not feasible for real time use. • Cannot find the exact solution but it finds best solution. • Initial guess biases the final output. • Genetic Algorithm (GA) is slow.

Among all conceivable organization decides that implant grammatical data, measurable standards, for example, MDL (Minimum Description Length) and Gibbs dissemination are being utilized so as to choose the ideal understanding, according to Geman et al (2004). This methodology is a starter which endeavors to consolidate measurements with language structure, however, sadly because of the eager compositionality process, it is unmanageable. In this paper, a novel procedure that looks to dissect and perceive basic segments inside an entire sorted out example is proposed.

Krogh et al (1994), and a digit number can be seen as a blend of strokes. The proposed structure utilizes a powerful multi-objective innate arrangement called Extreme Deep Learning Machine in the element extraction and determination strategy. Since EDLM decreases, the computational multifaceted nature of standard inherited figuring, it can find a predominant spread of courses of action and better association near the veritable Pareto optimal. Moreover, the suggested system used a capable summarized Multi shrouded Layer Feed-Forward Networks (MLFN) computation, called Extreme Deep Learning Machine (EDLM), as a gathering characterization.

2.7 SUMMARY

The guideline duty of this investigation is to fabricate another copper plate OCR structure that oversees division free pictures of Tamil words, by grasping both reality-based segment decision and EDLM in a Tamil OCR system in light of the fact that the system hopes to show up in any event affirmation botch, the tiniest running time, and the most direct structure.

CHAPTER 3

3. RESEARCH OBJECTIVE

3.1 THE MAIN OBJECTIVE OF THIS RESEARCH WORK

- To study the existing techniques for recognition of copper plate image Tamil characters.
- To classify ancient Tamil scripts written in digitized form of copper plate inscription scripts.
- To find appropriate preservation measures in order to improve the state of preservation of copper plate monuments and improve the clarity of copper plate Tamil character in the form of digital image.
- To preserve the valuable information that treasures up the intelligence of human kind that should be retained by a safe mode of recovery to revive them.
- To find ecologically approachable corrosion removal methods using nontoxic chemicals.
- To develop and apply the new framework combined segmentation technique and feature extraction on the handwritten copper plate images collected from different sources.
- To segment the text into individual characters.
- To find an optimized solution for classification of 11th century handwritten copper plate Tamil scripts.
- To improve the preprocessing technique by using proposed chemical solution.

- To enhance feature extraction techniques for improved version of copper plate digital image and to extract features from the characters that are size, font and slant invariant.
- To develop two different classification algorithms namely Extreme Deep Learning Machine and Complex Extreme Deep Learning Machine to achieve best solution for copper plate Tamil character recognition.

3.2 MAJOR CONTRIBUTIONS AND ACHIEVEMENTS

The main contributions are as follows:

- A detailed literature survey about segmentation, feature extraction and classification has been done.
- New algorithms have been proposed for line segmentation, half character segmentation, segmentation of touching modifier from consonant in middle region and lower modifier segmentation of handwritten copper plate Tamil character recognition.
- A new feature set has been developed for recognition of handwritten copper plate Tamil character recognition. Statistical features have been used for feature extraction.
- Different classifiers like ELM based EDLM and CEDLM classifiers have been proposed for classification and recognition.
- ELM is utilized in single-hidden layer feed-forward neural networks. It gives better performance than traditional tuning-based learning methods for feed-forward neural networks in terms of generalization and learning speed. It randomly chooses the parameters of hidden nodes and analytically determines the output weights.

Thus, the training is extremely fast and efficiently completed without time-consuming iterations.

- C-ELM achieves much lower symbol error rate (SER) and has faster learning speed than ELM. It does not recalculate the output weights of all the existing nodes when a new node is added.

3.3 METHODOLOGY

The proposed framework process is the cause for individuals to have a place in the various fields of work. This procedure includes catching the picture of copper plate Tamil Text and duplicating it in the checked picture. Examined Image needs to go through the proposed calculations which will assist the picture with getting changed over in advanced content configuration. For Character Mapping, the American Standard Code for Information Interchange (ASCII) is applied to Tamil characters and textual style is coordinated with its comparing format changed over which is then secured as standardized content translation dialects.

Training: For classification, a discretionary set of models for each letter is taken as the training set for that letter. Every case of this training set is iterated against the prepared set. The prepared set is the combination of a picture and a character. This iterative process is then run on all models in the preparation set and the resultant vector obtained from the picture is logged into a group of vectors. This bunch of vectors portrays the entirety of the models in the preparation set against the proposed character which is composed within the prearranged training set. Each bunch of vectors is dissected to locate a mean vector for the specific letter.

To recover the character of another picture we take the good ways from each mean vector to the information picture vector (as portrayed previously). A separation is figured to each mean-vector of the different letters. The letter that yields the littlest separation is utilized to order the picture. In the event that the littlest separation surpasses an edge, at that point the picture is announced to have a place outside the preparation set.

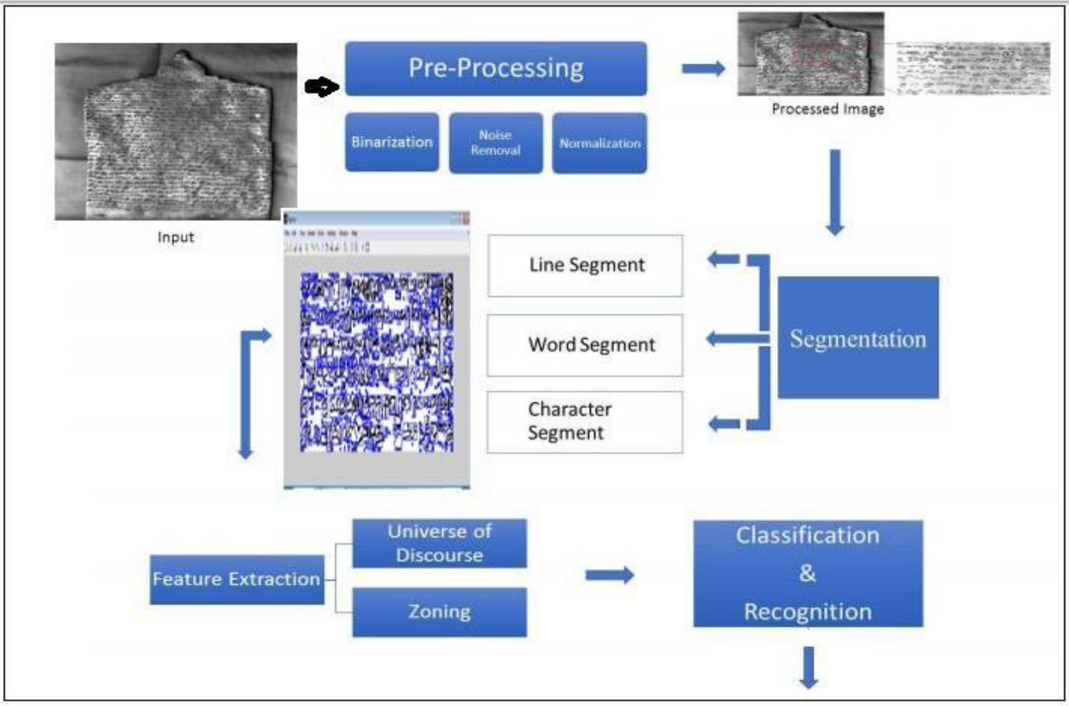


Figure 3.1: Tamil Character Recognition Methodology

3.4 SUMMARY

Based on proposed methodology, in this work we have implemented three different methodologies which focused on Eco-Friendly cleansing technique followed by Feature Extraction of Tamil handwritten font recognition.

CHAPTER 4

4. HISTORICAL METALLIC MONUMENTS PRESERVATION TECHNIQUES AGAINST DETERIORATION

4.1 INTRODUCTION

Past civilization’s information is acquired from Historical monuments. Researchers are concentrating on their research that focus on the cultural heritage preservation of historical manuscripts. Typically, these monuments are classified either within the scope of hand written documents, stone engravings, wooden engravings, metal plate inscriptions and mud or brick-based molds. There are various sources of historical copper monuments that provide us the overview of past history which is the prime focus of the researchers today as well as future followers. These copper monuments are slowly getting deteriorated because of environmental pollution, biological and anthropogenic activities. Thus, it is necessary to preserve these copper monuments from the aforementioned reasons. This analytical work elaborates on the common impacts of corrosion on copper plates and the technique adopted to retrieve the historical data from those plates.

Most of the identified degradation effects on historical copper monuments are mainly due to change in atmospheric exposure, sedimentation of dirt over time, biological contamination, etc. Critical impact on copper monuments is noticed widely on products

that are exposed to outdoor atmospheric environments. Climatic factors such as variations in temperature, comparative humidity, precipitation, snow and gaseous or hard pollutants released by industrial happenings are also main reasons for these deteriorations. Lesser impacts are noticed on copper monuments that are stored in indoor environments. Whereas, even in closed environments, prolong deposits or long-term storage impacts are seen which can cause damage to the stored copper objects.

4.2 CLASSIFICATION OF CORROSION

Deterioration of metals is defined as spontaneous destruction of metals in the course of their chemical, electrochemical or biochemical interactions with the environment. Based on the environment, corrosion is classified into,

1. Dry or Chemical Corrosion
2. Wet or Electrochemical Corrosion

4.2.1 Dry or Chemical corrosion

Direct chemical attack on metal by gases present in atmosphere triggers metal corrosions. Gases such as oxygen, halogen, hydrogen sulphide, sulphur dioxide, nitrogen or anhydrous inorganic liquid, etc., are the main root cause for these chemical reactions.

4.2.2 Wet or Electrochemical Corrosion

Electrochemical corrosion takes place due to:

i) The formation of oxide deposits on anodic and cathodic areas or parts in contact with these
ii) Presence of a conducting medium
iii) Corrosion of anodic areas
iv) Deposits of large number of tiny galvanic cells along with impurities and moisture

This involves floating of electron-current between the anodic and cathodic regions. At the anodic region oxidation takes place (liberation of free electron), hence anodic metal is destroyed via both dissolving and assuming blended state (including oxide, etc.). Hence, corrosion usually happens at anodic areas.

4.3 CORROSION IN COPPER

Corrosion products in copper layer are usually called patina. Surface layers formed because of corrosion of copper due to repetitive processes of dampening and drop-out of basic salts from saturated electrolytes at lowest pH levels. Development of patina is a long-term process which happens over a period of time in multiple stages. Different colored layers are formed due to the formation of different type of copper salts. Resultant colors in copper plate are formed because of chemical reactions over the plate. Table 4.1, given below shows the corrosion colors found on copper plates along with the information about their chemical compositions.

Table 4.1: Electrochemical Corrosion Color and Reason

Corrosion Color	Reason
Red and Brown	CU_2O cupreous oxide
Green	$CuCO_3$ Cu $(OH)_2$ Malachite
Blue	$2CuCO_3$ Cu $(OH)_2$ Azurite
Blue	Cu_2 $(CH_3COO)_2$ Basic copper acetate
Bluish Green	$CuCi_2$ cuprous chloride
Pale Green	CuCl(OH) Basic cupric chloride
Green Color	$CuSO_4$ 3Cu (OH_2) Basic copper sulphate

4.4 COMMON METHODS TO REMOVE CORROSION FROM COPPER

As stated by Knotkova et al (1994–1999), the general elimination of deterioration products can be divided into three groups;

1. Physical Methods
    a. Cleaning by Water under pressure
    b. Mechanical or abrasive cleaning (blasting),
2. Chemical cleaning Methods – 'draw-off' and pickling.
3. Other Cleaning Methods

4.4.1 CLEANING BY WATER UNDER PRESSURE

Removal of corrosion merchandise and dust via pressurized water is generally used alongside abrasive techniques or chemical cleansing. Pressures around 50 to 1000 psi are maintained for cleaning by water under pressure. Particles that get dropped off do not constantly stay on the floor and soluble components of corrosion crusts and layers on the surface are removed by using strain washing.

4.4.2 MECHANICAL OR ABRASIVE CLEANING (BLASTING)

Different materials used for mechanical cleaning or abrasive cleaning which include Polishing agents or pastes, mechanical brushes made up of metal wool or bristles, scalpel, special abrasive steel wool, etc. Cleaning using hand-operated mechanical equipment is tedious and time consuming. These techniques help to remove crusts, deposits and growth from a surface and keep a thin layer of patina.

4.4.3 CHEMICAL CLEANING

Drawing-off and pickling form the main methods of chemical cleaning did using different chemical solutions. The pickling process differs from drawing-off as it removes all layers of corrosion products down to pure metal. Drawing-off method can be carried out in workshops or even in the field. It is not feasible to achieve desired and uniform effect of agents on a sculptured surface. Corrosion products are removed only partially which might however help in future processes. Chelation or complex-forming

solutions or pastes are used for drawing off as per the brief note by Pago (1984):

- Without affecting the original metal, Alkaline rochelle salt (sodium potassium tartrate) is utilized for corrosion removal.
- Based on EDTA (ethylenediaminetetraacetic acid) on bronze adhesive or active corrosion application is used to remove corrosion, leaving them enriched with copper oxides allowing for forthcoming repagination.
- Treatment of ammonium hydroxide solution and concurrent cleaning with wool eradicates adhesive layers of corrosion products.
- Treatment of water and ammonium hydroxide (1:1) and simultaneous brushing with a metal brush or pumice removes thick incrustations.
- Treatment of a solution containing 1 volume part of powder soda, 5 volume parts of calcium hydroxide and 2 volume parts of sawdust removes corrosion products for strongly deteriorated and corroded surfaces.
- Immersion of corroded copper object in 5% Sodium sesquicarbonate.
- 5-15% Sodium Hexametaphosphate.
- Ethylene Diamine tetra acetate EDTA.
- 2-5% citric acid.
- Alkaline Rochelle salt.
- Alkaline glycerol.

4.4.4. Other Cleaning Methods

Alternate cleaning methods are usage of high-pressure steam or dry ice blasting using

little dry ice balls. Electrolytic techniques as well are used over Archaeological Copper and bronze objects to clean them. The main objective of the cleaning process is to eliminate destructive ion elements (chlorides) from the surface coverings. These methods help to retain the layers of patina and other components on a treated object. Limitations of existing cleaning listed below:

- ✓ Chemically inert material deposits cannot be removed.
- ✓ Fully plugged equipment will require mechanical cleaning since the circulation of chemical cleaning liquids would be impossible.
- ✓ Severe damage can occur if improper procedures are applied or unskilled people are employed in the application process.
- ✓ So physical and chemical cleaning is less effective.

4.5 ECO-FRIENDLY CLEANSING TECHNIQUE

In this work, we have introduced an eco-friendly phytochemical technique for the removal of corrosion merchandise from copper objects using Bryophyllum calcynium (Ranakalli plant) as the main course material for corrosion removal in conjunction with supplementary binding agents. The residues of this process are bio-degradable and hence do not harm the environment. The corrosion effect on the copper plates on using this composition is comparatively lesser with respect to other chemical treatments that are in practice currently.

Step 1: Take 100 grams of Bryophyllum calcynium leaves.

Step 2: Clean Bryophyllum calcynium leaves with water.

Step 3: Take 100 grams of raw rice, 5 grams of fenugreek seeds and 25 grams of Split Black-gram.

Step 4: Soak these ingredients for 3 hours in water.

Step 5: Drain the water and mix the Bryophyllum calcynium leaves.

Step 6: Grind these mixtures and make it as a paste.

Step 7: Allow this paste up to 6 hours for Fermentation process without adding any external chemical agents.

Step 8: Apply these fermented pastes over corroded metallic copper object.

Step 9: Leave this copper object along with binding agent for an hour.

Step 10: Remove the binding agent with clean water.

4.6 SUMMARY

The normal substances used to remove corrosion products from copper objects are also toxic which pollute the environment. So, it is essential to go with ecologically safer corrosion removal techniques using non-hazardous substances to prevent the impact on environment. Hence, it is necessary to conserve these copper monuments from deterioration and to find modern tools and techniques in the coming era to preserve this information from deterioration for the future researches. In this thesis, chapter 5, 6 and 7 proposes and develops feature extraction algorithm with classification algorithms and they also conduct experimental researches with proposed eco-friendly & cleaned copper plate images used in the proposed Tamil character recognition method to get better performance.

CHAPTER 5

5. EXTRACTION OF STATISTICAL FEATURES AND RECOGNITION OF CHARACTERS IN COPPER-PLATE INSCRIPTIONS

5.1 INTRODUCTION

Securing and restoring the old metal carvings helps to expand the knowledge on our history. They empower us to relate to people of different periods who followed different customs and had various inclinations. All out-of-date pieces are accessible in structures like stone and metal carvings, palm leaf replica and paper replica. Hundreds and thousands of reinvigorated old copper inscriptions contain critical information that are related to history. Statistical Feature Extraction method is used to examine the normal ways of obtaining information from copper plates with standard translation strategies. An intuitive device is proposed for epigraphers to examine and register old engravings in a helpful manner to recognize factors recorded in copper inscriptions. The system demonstrates promising outcomes to recover vital data from copper plates in an effective decipherable, less tedious helpful structure.

5.2 PROPOSED COPPER PLATE CHARACTER RECOGNITION ARCHITECTURE

CPCR (Copper Plate Character Recognition) is the first step towards Tamil character recognition from old copper plates. This incorporates photo analyzing of the subject by

each character, scrutinize the resultant picture by decoding each character from the picture into character codes, for example, ASCII characters regularly used in data representations. As shown in Figure 5.1, extensive works on present CPCR structures are combined with accompanying standard approaches to get pictures ready for desired outcome with respect to printed message which is also capable of getting impressive results from handwritten inscriptions that are transformed into pictures with poor affirmation quality.

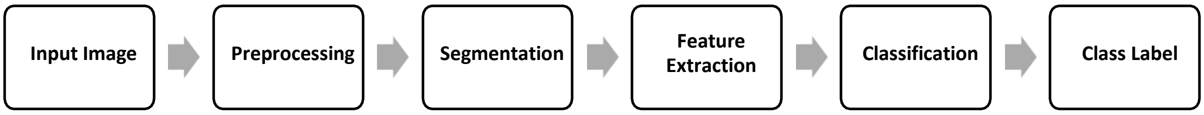


Figure 5.1: Block Diagram for Copper Plate Images Character Recognition

Given below are common issues that can surface while performing character recognition on manually written characters:

- Complexity of characters partition from foundation
- Nonstandard (interesting) type of images
- Nonlinear character area
- Different characters have distinctive inclination
- Neighboring images might get overlapped
- Some images might not be of uniform structure

Thus, this exercise is an essential mission than affirmation of standard printed content recognition. Generally, CPCR can be isolated into two segments: pre-preparing and emotionless character acknowledgment as shown in Figure 5.2. This research delineates the issue of pre-planning pictures with physically composed characters and the issues that arise during the progression of such trials. Eventually, this work has been developed into a Handwritten Text Detection model that has the capability to recognize even foggy, grainy and low-resolution pictures that were later moved into cumulative Data Extraction figurines. This helps to recover information in segregated plans and changes, reports into business-driven data that are better organized when dealing with examinations with limits. Shin et al (2001) stated to pick which keys to scan for and the substance affirmation computation that removes data from the aggregate of the chronicles that contain demonstrated keys paying little heed to where they are arranged inside the report. Before proceeding for material affirmation, the subjected image should be scrutinized well for dull zones or lighter areas to recognize each alphabet or digit.

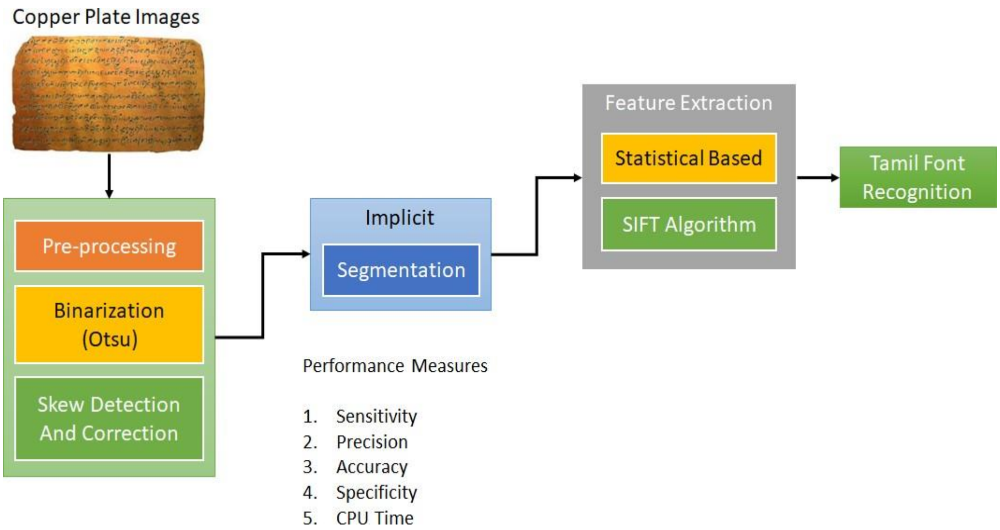


Figure 5.2: Proposed Architecture for Copper Plate Images Character Recognition

5.2.1. Otsu Image binarization

Binarization process relies on picture quality. Gayathri (2014) and Gayathri (2019) listed this circumstance based on picture quality.

- First procedure is to establish average using a histogram. Binarized pictures, adaptable with breakeven point figuring will give us the best results. The below mentioned points are well taken care of while dealing with pictures that consequent to resultant image.

- Generally, all photos can be used for additional pre-handling. In exceptional cases, combined tricks from numerical morphology are used (logical morphology is a speculation and strategy for the examination and treatment of geometrical structures). It could basically help in reproducing binarization characteristics of those photos. Changing parameters of morphological capacities with regards to pictures with moving degrees of significance will give us predominant results.

Pseudocode Steps involved are given below:

Step 1: Compute histogram and probabilities of each intensity level

Step 2: Setup initial $\omega_i(0)$ and $\mu_i(0)$
{where class probabilities ω and class means μ }

Step 3: Step through all possible thresholds t=1, …. maximum intensity

Step 3.1: Update ω_i and μ_i

Step 3.2:Compute $\sigma_b^2(t)$
where σ intra-class variance}

Step 4: Desired threshold corresponds to the maximum $\sigma_b^2(t)$

5.2.2 SKEW DETECTION AND CORRECTION

In the wake of completing the underlying advancement needed to get binarized pictures, get-togethers of white pixels made characters and dull pixels are used to form the establishment. Resultant data as explained by Martin (2004), Moreno (2009) and Voorhees (2005), clarify that there is a significant number of white pixels which are not part of characters that makes up the masses and hence create the disturbance that significantly impacts character affirmation computation.

Due to the credibility of transformation of the data picture while sifting and the effectiveness of many reported picture assessment systems in commotion of the image, the record inclination should be perceived and reconsidered. Projection profile is the most ordinarily used strategy to recognize the inclination.

5.2.3. SEGMENTATION

Segmentation is the route towards separating the subjected picture into content lines, words for a short time and later into characters. It is incredibly significant for gathering. In this research, we have actualized certain methodology to propose a strategy in diminishing the quantity of classes by character division and show that it brings about better character acknowledgment as depicted in Figure 5.3.

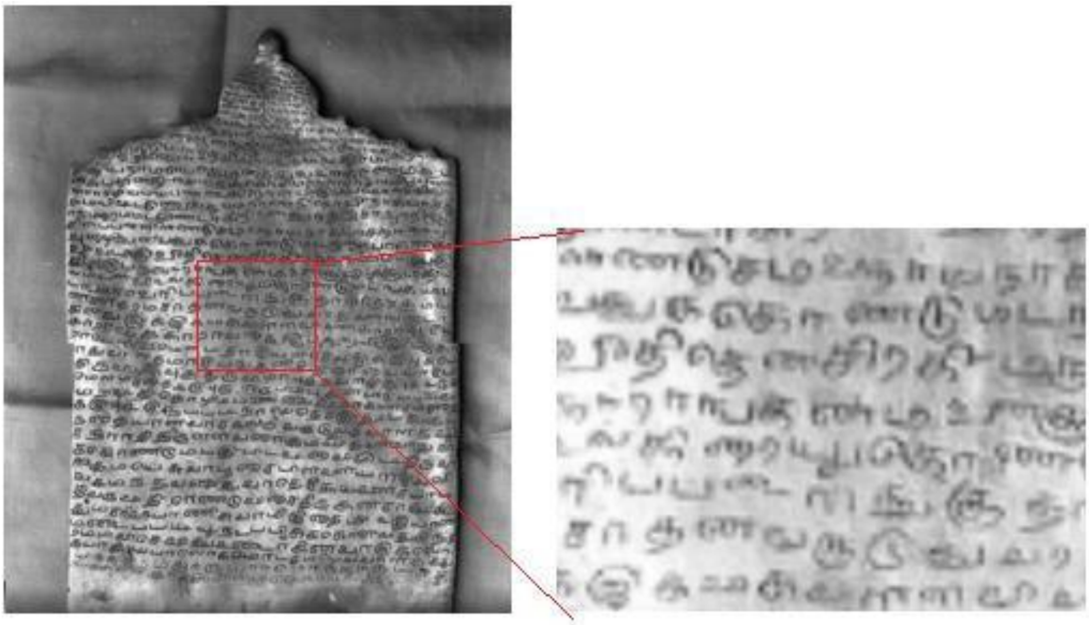


Figure. 5.3: Sample Processed Copper Plate Image

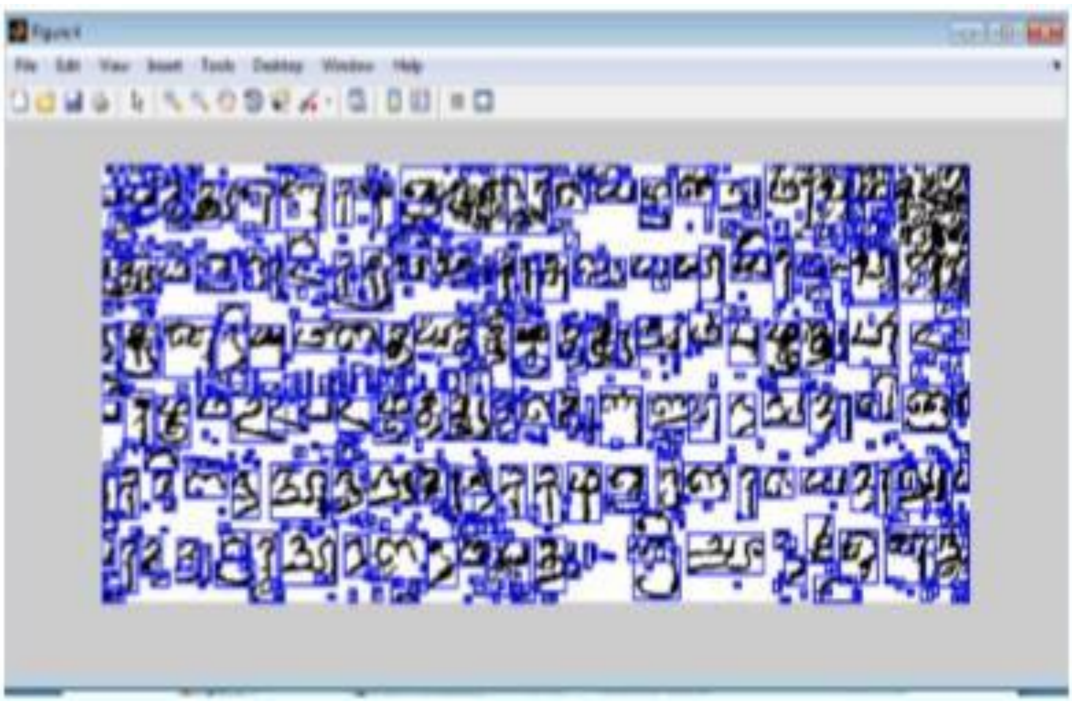


Figure 5.4: Copper Plate Image Segmentation

The character recognizer is a structural obstruct for freestyle penmanship acknowledgment and similar models can be utilized to perceive words as well. The words are perceived altogether without portioning them into letters. This is the best method which produces reasonable results just when the arrangement of potential words is little and known ahead of time, for example, the acknowledgment of bank cheques and postal location as shown in Figure 5.4.

5.2.4. FEATURE EXTRACTION

Highlights must be removed from the separated old Tamil subject in order to confine them into different classes. Feature extraction is portrayed as the path towards isolating the information from the rough data, which is commonly appropriate for gathering in the context of decreasing in-class plan uncertainties.

The highlights separated from this twofold picture are known as slope-based descriptors. Such highlights have been taken as valuables in transcribed content acknowledgment, human location and hand motion acknowledgments. The thought behind utilizing these highlights is that nearby shapes can be portrayed utilizing edge headings or by the dissemination of neighborhood angle powers without knowing the exact areas of the comparing slope focuses and edges. In Tamil textual style all the characters are of about a similar stature. Hence, we rescale the segregated character pictures to a standard elevation. We execute the calculation by first isolating this picture into vertical pieces of width w pixels.

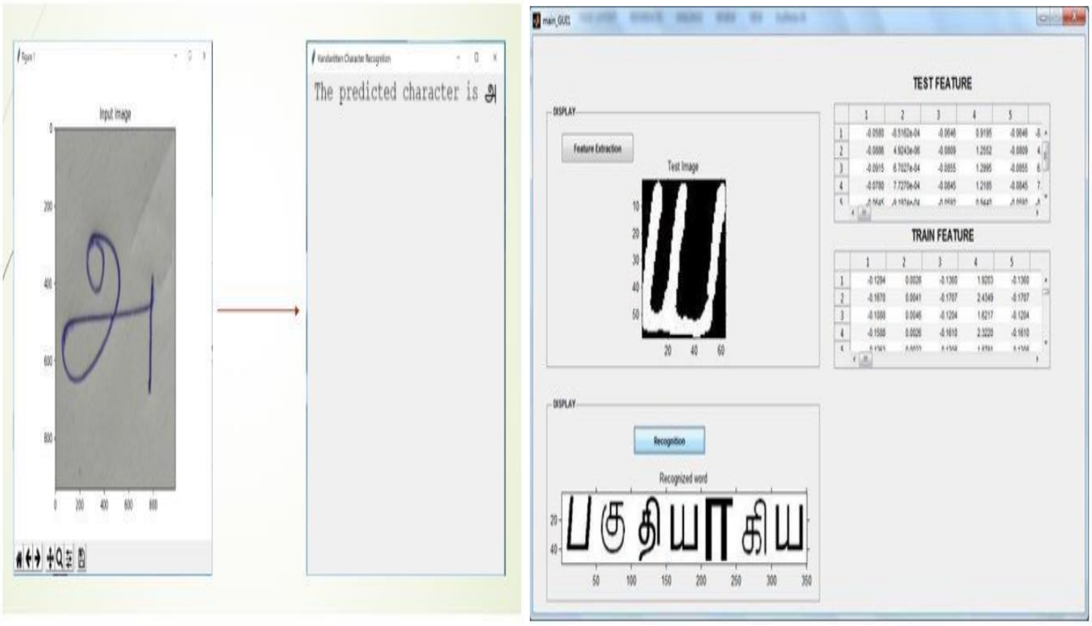


Figure. 5.5: Features Extracted from Cells Contained in Vertical Strips of Width W. Each Vertical Strip has H such Cells

We partitioned each strip into h locales which we call cells. In every cell, we register the histogram of slope headings over the pixels of the cell. These bearings are specified by direction canisters equally dispersed over 00 to 3600. The joined histogram routes us to the structure of the element. The parameters w and rabbit are fixed by approval, Gayathri (2014). We found that 5 containers are sufficient for a paired picture. Every pixel in a cell represents a weight being added to one of the histogram channels. A fixed weight of one demonstrating presence in that specific channel is utilized. We compute the angles utilizing a straightforward [−1 0 1] veil in both the X and Y bearings with no Gaussian smoothing. We found no improvement by utilizing subsidiary of Gaussian (DoG) portions to ascertain the slope.

The last and the hardest development are on character distinguishing proof. It will try to outline the character without jumping too much profound into nuances. It should call masses of white pixels which are a bit of a character which is our area of interest, ROI. Consequently, to complete all the past advances it got some pre-arranged pictures with compensations. For example, a couple of characters can be apportioned into parts or merged. The highlights are removed uniquely from the locale of the picture which contains closer view pixels.

In the photos underneath you can perceive how an issue with separating single character was settled:

- Resolving the issue with combined characters from transcribed content is harder. Since, composed characters have truly factored inclination, underneath you can see the after effects of character division calculations. Not all characters were isolated effectively in this pre-handling stage on the grounds that the product cannot perceive what it is partitioning.
- And the last advance in character identification is character ROI partition and cleaning.

5.3 EXPERIMENTAL RESULTS

The whole process of character identifying methodology is carried on various Copper plate images inscriptions sourced from different parts of Tamil Nadu, India. Copper plate inscriptions sourced in various periods depict the styles that were followed over time with respect to the types of stones used, polishing methods adopted, composition of the texts used, colors usages over copper plates, engraving techniques used on the copper plates as well as the locations chosen to erect the copper the plates.

Figure 5.6:a **Figure 5.6:b** **Figure 5.6:c**

Figure 5.6: d

Figure 5.6:e

Figure 5.6 a, b, c, d , e: Sample Copper Plate images

Many of these inscriptions are deteriorated beyond a limit that it becomes very difficult to acquire the vital data from those, especially when the surface is found to be impacted with corrosion or etching. Impacted over centuries, these deteriorations are beyond certain limit that the texts on them are in very poor condition and most of the connecting parts are already missing. The impact is found to be to such an extent that either some fragments are gone missing or there are sections that are no longer identifiable and recoverable. The performance result of the character spotting process on such images is also reported in this section.

Table 5.1: Desktop Performance (in %) and Average CPU Time (per spotting per template) for the Inscription Images.

Performance Measure	Sample Copper Plate Images				
	Fig 5.6 a As “a”	Fig 5.6 b As “b”	Fig 5.6 c As “c”	Fig 5.6 d As “d”	Fig 5.6 e As “e”
	Tamil Font Recognition				Symbols
Sensitivity	85.23	73.24	65.37	52.34	47.89
Specificity	99.14	85.67	84.89	92.45	92.54
Precision	78.90	43.56	32.78	51.67	47.78
Accuracy	89.12	84.23	84.35	65.78	90.37
CPU time (sec.)	01.11	00.78	00.76	00.75	00.65

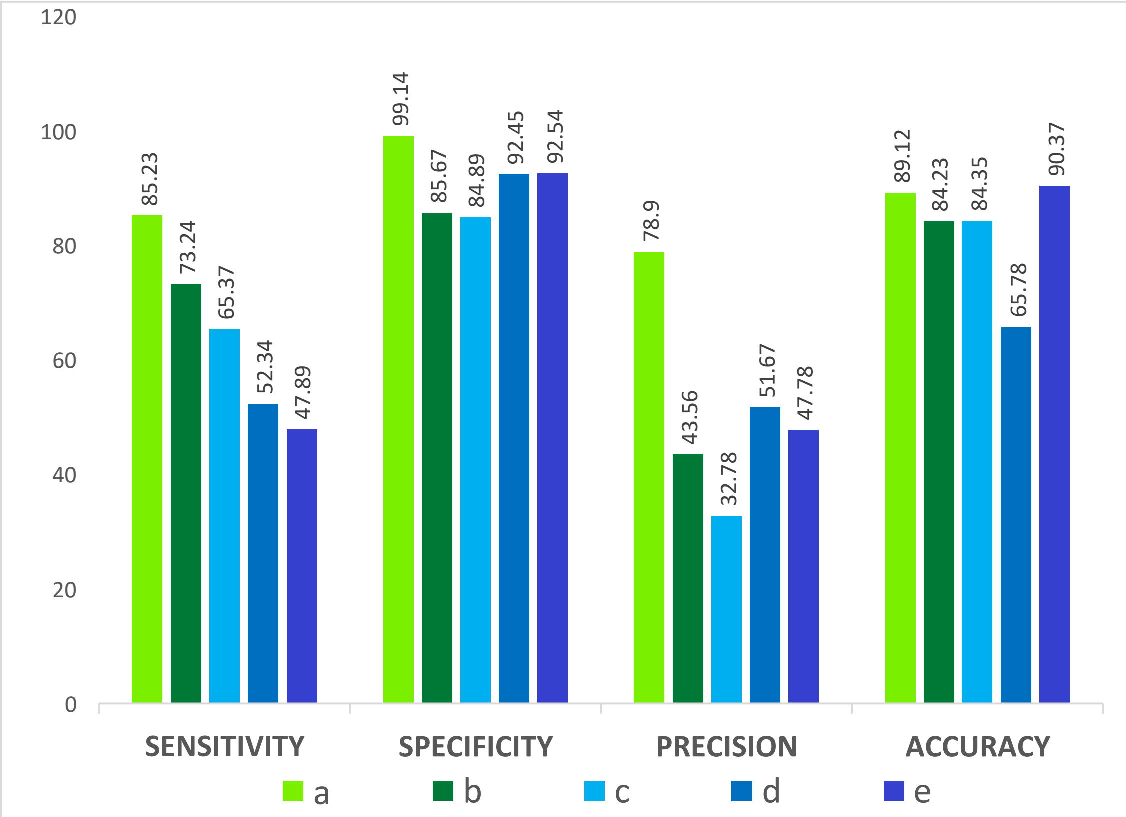


Figure. 5.7. Performance Measure for Different Sample Images

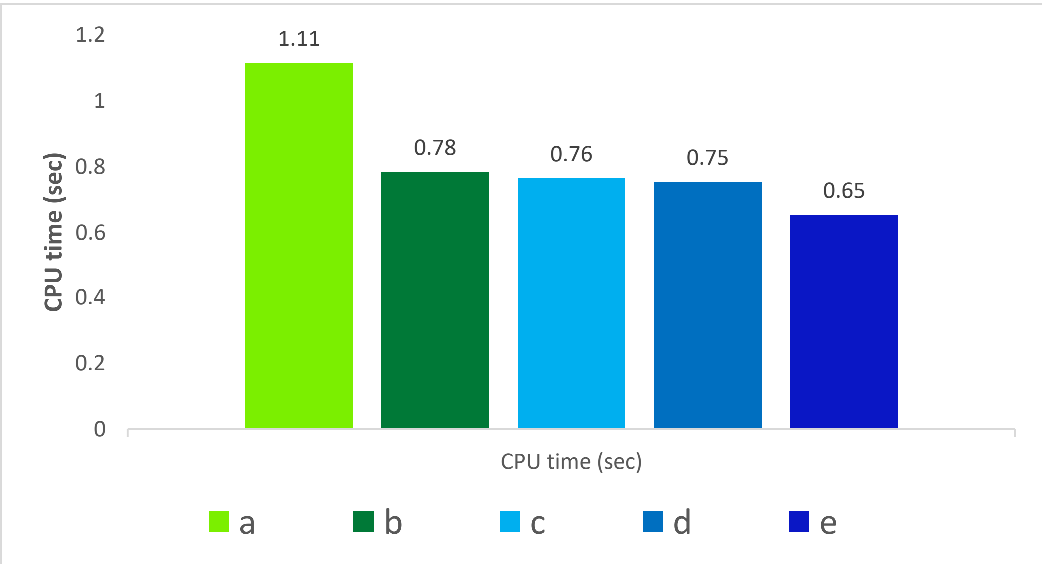


Figure 5.8: Difference Copper Plate Images Performance Measure: CPU time for Different Images

The enactment of the noticing technique is analyzed by approximating the events of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). For the spotting result of each character, we calculate the number of true positives (TP), false positives (FP), true negatives (TN), and false positives (FN), which correspond to correctly spotted, correctly rejected, incorrectly spotted and incorrectly rejected characters, respectively,

Sensitivity (True Positive Rate) = TP/ (TP+FN) ....(5.1)

Specificity (True Negative Rate) = TN/(FP+TN)...(5.2)

Precision = TP/(TP+FP)...(5.3)

Accuracy = (TP + TN) /((TP+FN+TN+FP)...(5.4)

5.4 SUMMARY

Copper PCR of hand written ancient inscriptions are fairly complex in interpretation. The present programming and created calculations are unable to accomplish 100% precision in retrieval. These CPCR can be expanded to be able to give an easily understandable and conclusive outcome by repeated pre-processing - acknowledgment grouping. In this research, it is proposed to build an interactive tool for epigraphers to read and archive ancient inscriptions in a convenient way. This replaces the tedious task of obtaining the estampages (exact replica of an inscription that cannot be transported) from copper plate inscriptions with ink-smeared manual dabbers, adopted in conventional practice. The proposed character spotting results are useful in creating a dataset of various characters of the concerned language and it can be helpful in studying the hierarchy of evolution of the scripts. Further, adopting such similar data patterns repeatedly is useful in training classifiers during recognition process. This is a probable candidate for future works.

CHAPTER 6

6. EXTREME DEEP LEARNING MACHINE (EDLM)

6.1 INTRODUCTION

Tamil which is conceivably one of the ancient spoken and ancient written languages in the world is the primary language in Tamil Nadu, South India. The primary wellspring of statistics about history is the stone engravings. Characteristic extraction technique is utilized to decide on an appropriate selection algorithm and their types of techniques to realize higher recognition precision that requires low computational overhead.

This study proposes a new strategy termed as Extreme Deep Learning Machine (EDLM), depicting a set of rules compatible for a class of machine learning that has a quick processing time. Also, the EDLM is devoid of certain demerits observed in gradient-based mastering strategies that implement epochs in search of nearby local minimal. The EDLM ruled-independent elucidations express variety of protection mechanism. Comparison with the experimental effects of other methodologies revealed the talent of the proposed system and confirmed enabling required feature choices.

6.2 PROPOSED EXTREME DEEP LEARNING MACHINE SYSTEM

As stated by Shin (2001), EDLM utilizes five procedures which are image acquisition, pre-processing, segmentation, feature extraction, selection and finally classification as shown in

Figure 6.1. The square graph shortens the fundamental segments of the proposed Tamil OCR framework model. The framework uses factual based element choice calculation to choose the ideal highlights and Extreme Deep Learning Machine (EDLM) classifier for acknowledgment of the Tamil OCR framework. The framework comprises of two primary similar stages: training phase and testing phase. The two stages incorporate five procedures executed successively in each preparation stage and testing stage as shown in Figure 6.2.

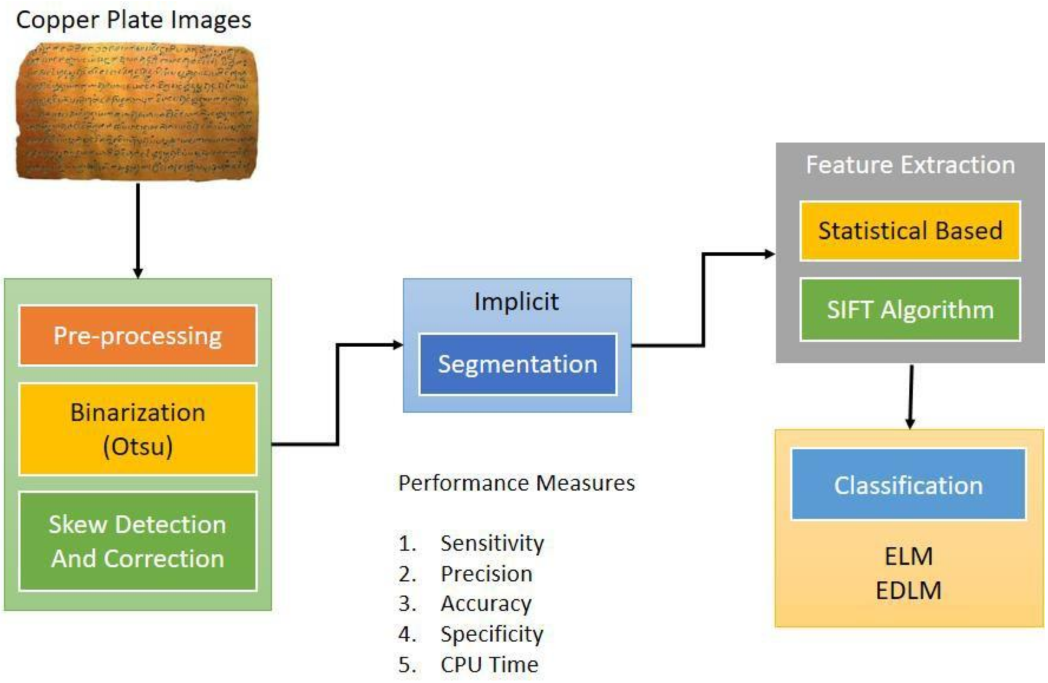


Figure 6.1: Proposed EDLM Architecture for Copper Plate Images Character Recognition

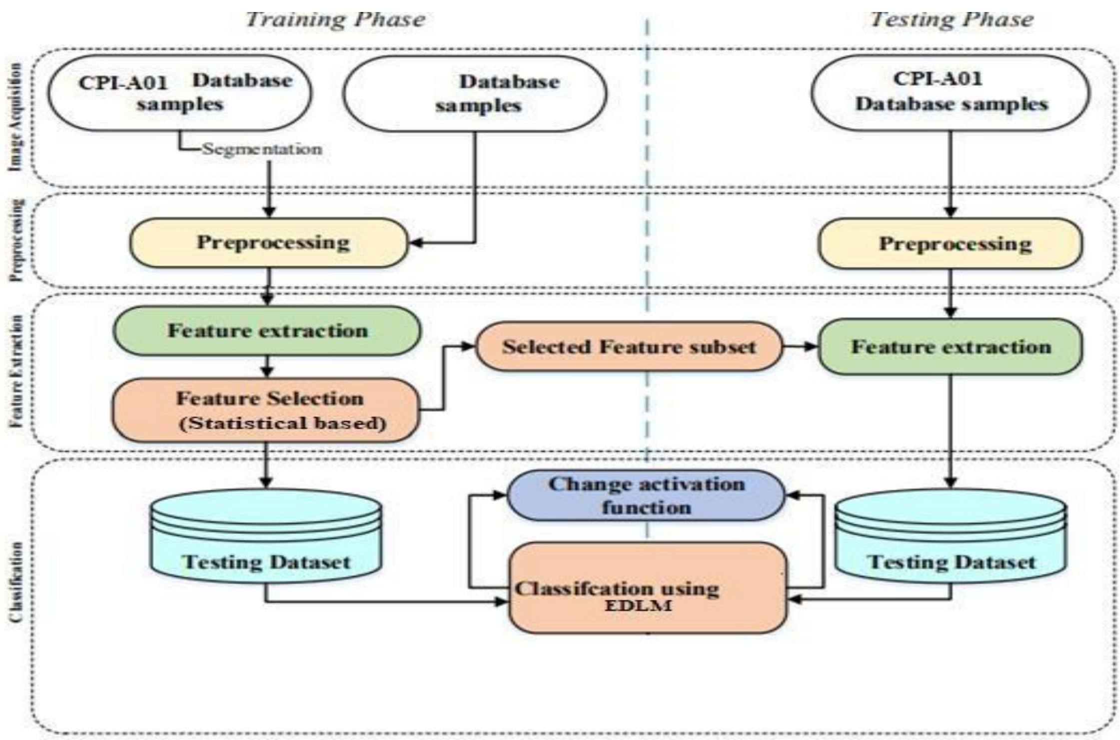


Figure 6.2: Detailed EDLM Architecture for Copper Plate Images Character Recognition

Pseudo code:

Step 1: Initialize numerical values between 0 to 1

Step 2: Set input weights aij and the bias of the hidden layer bj

Step 3: Calculate the output matrix H.

Step 4: Calculate the output weights V

$$V=H^{\dagger}T$$

Where $H^{\dagger}$ represents the generalized inverse matrix of the output matrix H.

6.2.1 Otsu Image Binarization

Thresholding is a fundamental system in image segmentation applications. Otsu (1979) method is considered worldwide threshold in which it depends on dark estimation of the picture by Bowyer (2001) and Shin (2001).

Otsu's method is characterized by its nonparametric and unsupervised nature of threshold selection and has the following desirable advantages.

- Simple procedure is used; Considers only zeroth and first order cumulative moments of the gray-level histogram.
- Multi-thresholding problem is handled by extending the virtue of the criterion on which the method is based.
- A stable and optimal threshold is selected automatically based on integration (global property) of the histogram and not on differentiation (local property).
- Extended analysis can be done (e.g., estimation of class Mean levels, evaluation of class separability, etc.).
- This binarization method is general which covers a wide scope of unsupervised decision procedure.

The essential recommendation of threshold is to settle on a best dark level edge on incentive for isolating objects of enthusiasm for a picture from the foundation dependent on their dim level dissemination. The dark level histogram of a picture is commonly considered to be composed apparatuses for development of thresholding calculations. By turning all pixels beneath some limit to zero and all pixels about that

edge to one, thresholding makes twofold picture. In the event, g (x, y) is an edge record of f (x, y) at different global threshold edges T, it very well may be characterized as g(x, y) = 1 if f(x, y) ≥ T = 0 in any case, g(x, y) is a threshold version of f(x, y) at some global threshold T.

Thresholding can be categorized into two classes: global threshold and local (adaptive) threshold.

- In the global threshold, a single threshold value is used in the whole image. When T depends only on f (x, y), only on gray-level values and the value of T solely relates to the character of pixels, otherwise.
- In the local threshold, a threshold value is assigned to each pixel to determine whether it belongs to the foreground or the background pixel, using local information around the pixel.

6.2.2 Skew Detection and Correction

Manually written copper plate composition may at first be slanted or skewness may be present in copper content checking process. This impact is inadvertent in numerous genuine cases, and it ought to be killed in light of the fact that it successfully diminishes the exactness of the sequential procedures, for example, division. As elaborated by Martin (2004), Moreno (2009), Voorhees (2005), Gayathri (2014), Skewness is remediated by utilizing projection profile analysis. A twofold picture converted into one-dimensional exhibit (projection profile) is known as projection. Each line in projection profile has a value that produces various dark pixels in relating columns of the picture and lines on record are spoken to as level histogram profile. For those

pictures that contain zero slanted edge, the flat projection profile has a channel which is equivalent with the space between the lines. And furthermore, the most extreme pinnacle tallness which is equivalent to content lines stature are present in archived pictures. In this way, this strategy computes the distinction in projection profile at various divergent edges in equivalent points that has the most contrast pattern.

6.2.3 SEGMENTATION

Segmentation is the route towards separating the reported picture into content lines, words and subsequently into characters. It is incredibly important for gathering reason. Right now, Mallikarjunaswamy (2001), Olarik et al (2008) and Sagar et al (2008) proposed a verifiable methodology and strategy to decrease the quantity of classes by character division and to show that it brings about better character acknowledgment. The character recognizer is a structure which observes that the freestyle penmanship acknowledgments in similar models can be utilized to perceive words. The words are perceived totally without dividing them into letters. This is best feasible just when the arrangement of potential words is a little and known ahead of time, for example, the acknowledgment of bank cheques and post allocation.

6.2.4 FEATURE EXTRACTION

The highlights removed from parallel pictures are measurable descriptors. Such highlights have been seen valuable in manually written content acknowledgment, which are the human discovery and hand-motion acknowledgment. The thought behind utilizing these highlights is that nearby shapes can be portrayed utilizing edge headings

or by the dissemination of neighborhood slope forces without knowing the exact areas of the relating inclination focuses and edges. In Tamil text style, all the characters are of about a similar stature. Thus, Vikas (2011) and Gayathri (2019) rescale the detached character pictures to a standard tallness. We actualize the calculation by first partitioning this picture into vertical portions of width w pixels.

6.2.5 CLASSIFICATION

11th century hand written Tamil contents are regularly assembled into four classes, to be specific, Vowels, Consonants, Composite characters and Aydham. These four classes are taken for characterization reasoning right now. Customary calculations are far slower than required in light of the fact that the slope-based learning calculation and the parameters must be tuned iteratively. In this way, Extreme Deep Learning Machine (EDLM) is utilized for arrangement.

Feature extraction and selection procedures produce the element vector utilized in the arrangement organize. Characterization is the dynamic procedure in the OCR framework that utilizes the highlights removed from the past stages. The characterization and calculation by Rafael et al (2007), Lu et al (2003) and Anupama et al (2013) is educated with the preparation dataset, which at that point is encouraged with the testing dataset to perceive the various classes (each class is a word). Accomplishing a high acknowledgment rate requires a ground-breaking order system that beats its counterpart methods as far as speed, straightforwardness and acknowledgment rate is considered.

The proposed framework by Shin (2001) uses EDLM, a quick and effective learning calculation, characterized as a summed up Multi concealed Layer Feed forward Network (MLFN) as shown in Figure 6.3. Basics of ELM methods are made out of twofold: inclusive of all guess capacity with irregular concealed layer and different learning strategies with simple and quick usage

$$f_L(x) = \sum_{i=1}^{L} \beta_i g_i(x) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x),\ x \in R^d, \beta_i \in R^m) \quad \text{..... (6.1)}$$

Where a_i and b_i are the learning parameters of hidden nodes and β_i is the weight connecting the i^{th} hidden node to the output node, $\beta_i G(a_i, b_i, x)$ is the output of the i^{th} hidden node with respect to the input x as stated by Martin(2004).

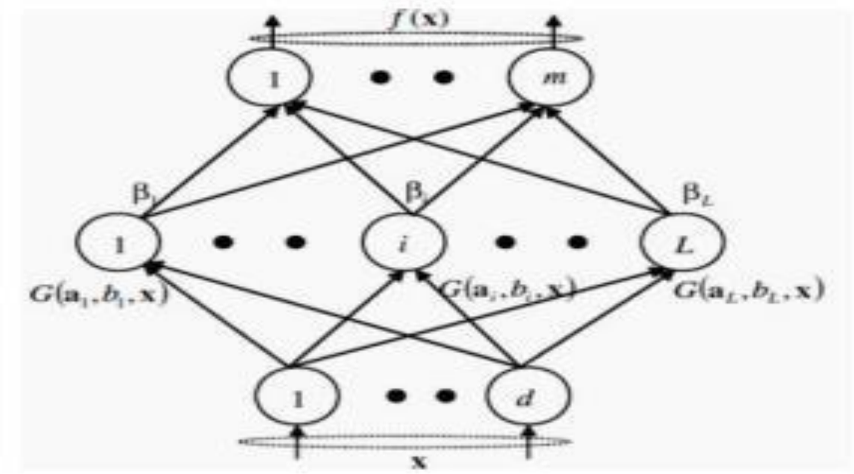


Figure 6.3: ELM Feed Forward Network Architecture

For N arbitrary samples $(x_i, t_i) \in R^d \times R^m$, the MLFN with L hidden nodes is modeled as:

$$\sum_{i=1}^{L} \beta_i G(a_i, b_i, x_j) = t_j,\ j = 1, ..., N \quad \text{........ (6.2)}$$

The above equation can be written compactly

$$H\beta = T$$

$$H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} G(a_1, b_1, x_1) & \cdots & G(a_L, b_L, x_1) \\ \vdots & \cdots & \vdots \\ G(a_1, b_1, x_N) & \cdots & G(a_L, b_L, x_N) \end{bmatrix}_{N \times L} \quad \text{........ (6.3)}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m}, \text{ and } \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \quad \text{........ (6.4)}$$

H is called the hidden layer output matrix of the MLFN and T is called target labels. The essence of EDLM tends to minimize ||H T- β|| and ||β|| , so the most extreme number of shrouded hubs required isn't bigger than the quantity of preparing tests. We have created Extremely Deep Learning Machine with Multi Layered Forward Network including measurable element determination for getting higher precision and quick calculation.

The performance of ELM is contrasted with Probabilistic Neural Network (PNN), in which the operation is organized into multilayered feed-forward network with four layers namely input layer (data set), pattern layer (trained data), summation (iterated data) to achieve result and out layer (required character recognition) and it is seen that 70.19% and 78.73% of exactness is accomplished by PNN and ELM individually. So as to expand the exactness of order further, EDLM is utilized. The exhibition of EDLM is estimated by contrasting it on ELM. EDLM noticed to be providing a grouping precision of 80.30% when contrasted with ELM. So as to build the exactness, we lessen the number of concealed neurons utilized and diminish the time taken for preparing an EDLM which has been proposed to be used in this research work.

6.3 EXPERIMENTAL RESULTS

To check the productivity and legitimacy of the proposed framework, the framework was tried for precision and the outcomes were contrasted with that of aftereffects of past frameworks on similar databases.

So as to ensure the nature of pictures, all pictures were gathered from continuous copper plate pictures taken by us from different sources. Preparing sets contained excess of 70 preparing pictures and in excess of 50 tests from CPI-A01 by Olarik (2008). The individual content lines of CPI-A01 database were portioned physically to isolate them into words. Instructional courses were 25 distinctive Tamil words in various sizes, directions, clamor degrees and textual styles. The examinations were led on an AMD Quad-centered, 2 GHz processor, 4 GB DDR3 Ram PC and Windows 8.1 working framework. The code was written in MATLAB language utilizing MATLAB 2011Rb programming. The proposed Copper plate character spotting technique is tried on the pictures of various copper plate engravings gathered from different places of Tamil Nadu, India. Copper plate engravings made up of different lines that are present allover Tamil Nadu show extraordinary highlights in their style concerning the type of stones used, cleaning done, piece of content used, recording method on copper plate with shading, etching the content on copper plate and furthermore, the circumstances of raising the copper plates from suitable spots. A considerable lot of these engravings are disintegrated so gravely that it is hard to distinguish the important information, especially when the surface is in depletion or when it is carved. Because of hundreds of years of decaying effect, dominant parts of these antiquated writings are in poor

condition and numerous content bits are beyond recognition. The harm has happened to such a degree, that neither the pieces exist nor the segments are again noticeable and recuperates. The presentation aftereffect of the character spotting process on such pictures is likewise announced now.

.







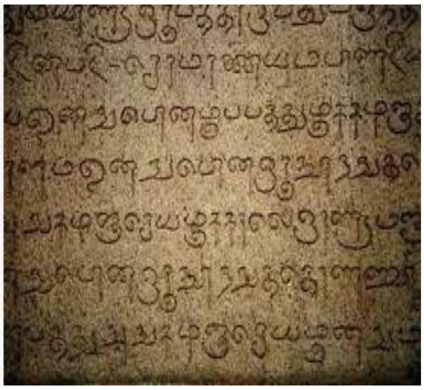



Figure. 6.4: Sample Copper Plate images

Figure 6.5: Desktop Performance (In %) and Average CPU Time (Per Spotting Per Template) for Inscription Images

At first, the element dataset contains full 14 highlights. The preparation dataset is of measurement 102x14 and the testing dataset is of measurement 22x14. These two sets are applied to EDLM arrangement. In the principal exploration, ELM is applied with various actuation capacities: Sigmoidal, Sine, Hardlim, Linear, Triangular, and Radial premise. An activation function is any nonzero function used to transform the activation level of a neuron into an output signal.

The beginning number of concealed neurons was self-assertively picked to be 50, as 50 is practically 50% of the preparation dataset. Be that as it may, numerous criteria were utilized in assessment of the framework, which are preparation time as characterized as the time spent on preparing ELM, testing time which is the time spent on anticipating all testing information and preparing/testing exactness which is the root mean square of right characterization.

Table 6.1. EDLM Applied with Different Activation Functions

S.NO	Activation Function	Training Time	Testing Time	Training Accuracy	Testing Accuracy
1	Sigmoidal	0.00961s	0.00012s	0.4554	0.4091
2	Sine	0.00521s	0.06250s	0.9703	0.9091
3	Hardlim	0.00761s	0.06250s	0.2376	0.2376
4	Triangular	0.23440s	0.00000s	0.1584	0.1364
5	Radial basis	0.15630s	0.00014s	0.2376	0.3182

Table 6.2. Comparison between the Proposed System Classifier and Different Classifiers on the CPI −A01 Database

S.NO.	System Classifier	MaxTest Accuracy
1	Probabilistic Neural Network (PNN)	70.18%
2	Support Vector Machine (SVM)	**72.4%**
3	Extreme Learning Machine (ELM)	**78.13%**
4	Extreme Deep Learning Machine (EDLM) Proposed System	**85.87%**

In addition, the classifier framework in Rafael et al (2007) relies upon some heuristic punishments and division procedures that incredibly influence the Sift descriptor exactness. The targeted framework is designed excluding component descriptors and it utilizes interpretation and scale invariant highlights.

6.4 SUMMARY

Copper Plate Optical Character Recognition (CPOCR) for composed content is an exclusively testing and open territory of research area. This research on copper plate Tamil OCR for manually written words is built dependent on a blend of Extreme Deep Learning Machine (EDLM) classifier with Multi-hidden Layer Feed Forward Network and factual based component determination.

In the beginning, the framework utilized a 14 highlights dataset. At that point, information was sustained into EDLM organizer which was a quick and basic multi concealed layer in feed forward system (MLFN). The framework accomplished high acknowledgment precision of 85.87% for various examples in an extremely brief timeframe. EDLM stays away from the nearby least snares and long preparation time of conventional neural systems and the concealed layer of MLFNs also need not be tuned. In addition, Statistics based element determination chooses the most characterizing highlights that diminishes datasets multidimensional nature by 57% and improves the exhibition fundamentally.

CHAPTER 7

7. COMPLEX EXTREME DEEP LEARNING MACHINE (CEDLM)

7.1. INTRODUCTION

Recognizing ancient Tamil characters enable archaeologists to reveal historical events in Cholas period that dates back to 12th century which reduces huge efforts for the archaeological experts. The future researches in the field of archaeology will have negative impact due to inefficiency in the manual procedures. Optical Character Recognition (OCR) functionality is used to recognize ancient Tamil Inscriptions. OCR module of application is mainly focused in this research. In this, we propose a Complex extreme Deep learning machine algorithm (CEDLM) that adds some hidden layers to the original ELM network structure which randomly initializes the weights between the first hidden layer and the input layer as well as the bias of the first hidden layer, utilizes the method (make the actual each hidden layer output approach the expected hidden layer output) to calculate the parameters of the hidden layers (except the first hidden layer) and finally uses the least square method to calculate the output weights of the network.

The subsequent calculation is a feature extraction (Scale Invariant Feature Transform (SIFT) algorithm to detect and describe local features in images. Comparison with the experimental results of other methodologies revealed the proficiency of the proposed

system and demonstrated that the feature selection approach increased the accuracy of the classification process.

7.2. PROPOSED COMPLEX EXTREME DEEP LEARNING MACHINE SYSTEM

The architecture which is represented in Figure 7.1, shows the square graph that abridges the fundamental segments of the proposed Tamil OCR framework model. The framework uses factual based element choice calculation to choose the ideal highlights and Complex Extreme Deep Learning Machine (EDLM) classifier for acknowledgment of the Tamil OCR framework. The framework comprises of two primary similar stages: training phase and testing phase. The two stages incorporate five procedures executed successively in each preparation stage and testing stage.

Pseudo code of CEDLM

- Adjust the structure of ELM neural network
- Consider first hidden layer as the first hidden layer, where as
- The second hidden and the third hidden layer together as one hidden layer.
- Consider new structure of the network is the same as the two hidden layer ELM network Calculate weights matrix βnew between the second hidden layer and the output layer.
- Repeat the calculation based on actual samples and calculate the weights β,
- To improve the generalization ability of the network separates merged three hidden layers, so that the structure has three hidden layers.
- Calculate expected output of the third hidden layer

The five procedures are image acquisition, pre-processing, segmentation, feature extraction and selection and finally classification as shown in Figure 7.2.

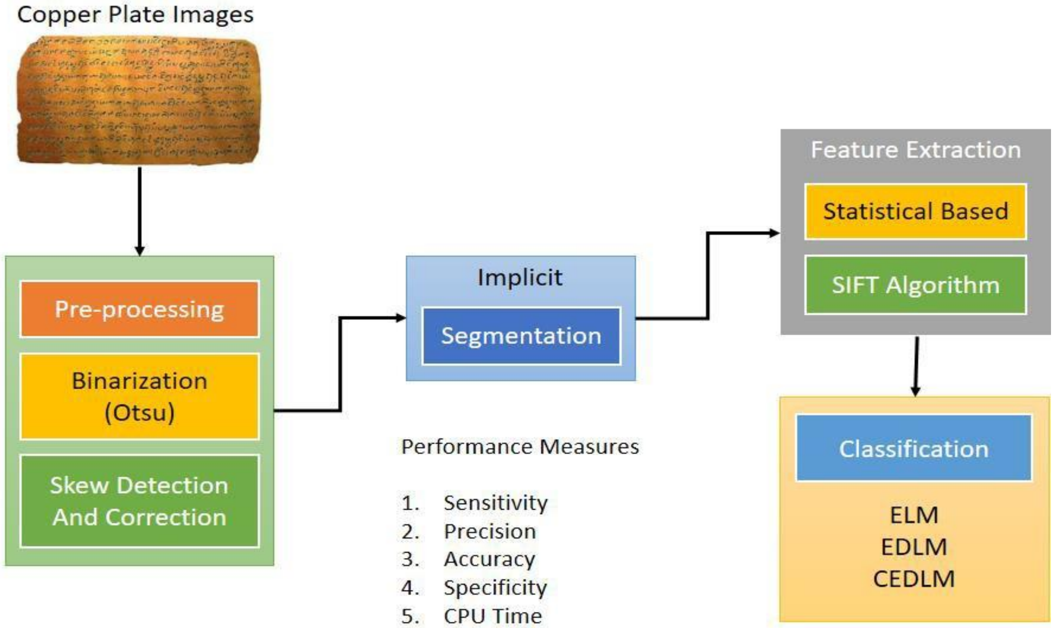


Figure 7.1: Proposed CEDLM Architecture for Copper Plate Images Character Recognition

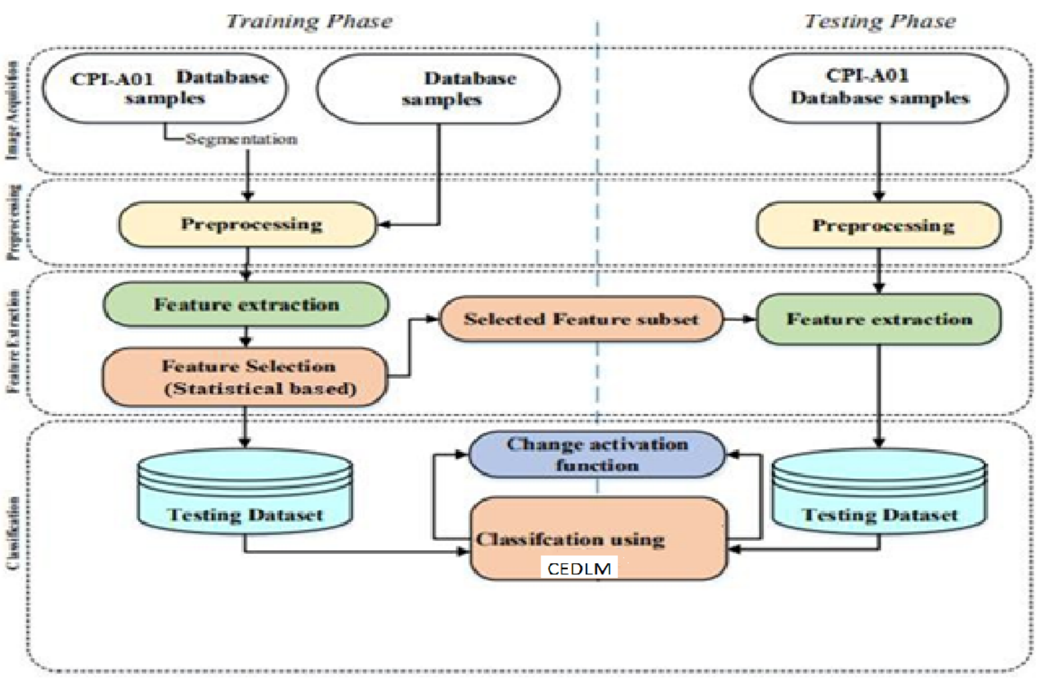


Figure 7.2: Detailed CEDLM Architecture for Copper Plate Images Character Recognition

7.2.1 Otsu Image Binarization

Thresholding is a fundamental system in image segmentation applications. Otsu method is somewhat worldwide thresholding in which it depends on dark estimation of the picture, with reference to Bower et al (2001) and Shin et al (2001). The essential recommendation of thresholding is to settle on a best dark level edge on incentive for isolating objects of enthusiasm for a picture from the foundation dependent on their dim level dissemination. The dark level histogram of a picture is commonly considered to be composed apparatuses for development of thresholding calculations. By turning all pixels beneath some limit to zero and all pixels about that edge to one, thresholding makes twofold picture. In the event that g (x, y) is an edge record of f (x, y) at different global threshold edges T, it very well may be characterized as $g(x,y) = 1$ if $f(x, y) \geq T = 0$ in any case.

7.2.2 Skew Detection and Correction

Manually written copper plate composition may at first be slanted or skewness may be present in copper content checking process. This impact is inadvertent in numerous genuine cases, and it ought to be killed in light of the fact that it successfully diminishes the exactness of the sequential procedures, for example, division. Skewness is remedied by utilizing projection profile Analysis proposed by Martin et al (2004), Moreno (2009), Voorhees (2005) and Gayathri (2014) [11-15]. A twofold picture into one-dimensional exhibit (projection profile) change is known as projection. Each line in projection profile has a worth that produce various dark pixels in the relating column of the picture and lines on record are spoken to as level histogram profile. For those pictures that contain

zero slanted edges, the flat projection profile has channels which are equivalent with the spaces between the lines. Furthermore, the most extreme pinnacle tallness is equivalent to content lines stature present in archive pictures. In this way, this strategy computes the distinction in projection profile at various divergent edge equivalents to points that have the most contrast areas.

7.2.3. Segmentation

Segmentation is the route towards separating the subjected picture into content lines, words and subsequently into characters. It is incredibly important for gathering results. Specifically, the proposals by Mallikarjunaswamy (2011), Olarik (2008) and Sagar (2008) towards verifiable methodology and strategy in decreasing the quantity of classes by character division show that it brings better character acknowledgment. The character recognizer is a structure renderer for freestyle penmanship acknowledgment since similar models can be utilized to perceive words. The words are perceived totally without dividing them into letters. This is feasible just when the arrangement of potential words is a little and known ahead of time, for example, the acknowledgment of bank cheques and postal location.

7.2.4. Feature Extraction

The highlights removed from parallel pictures are measurable descriptors. Such highlights have been seen as valuables in manually written content acknowledgment, human discovery and hand motion acknowledgment. The thought behind utilizing these highlights is that nearby shapes can be portrayed utilizing edge headings or by the

dissemination of neighborhood slope forces without knowing the exact areas of the relating inclination focuses and edges. In Tamil text style, all the characters are of about a similar stature, according to Gayathri (2019) and Vikas (2011). Thus, we rescale the detached character pictures to a standard tallness. We actualize the calculation by first partitioning this picture into vertical portions of width w pixels.

Scale Invariant Feature Transform (SIFT) Algorithm SIFT Huttenlocher (1993) Moreno (2009) is the component extraction calculation to recognize and depict nearby highlights in pictures. The key focuses are first separated from the arrangement of reference picture and put in a database. An engraving test is perceived by independently looking at each component from the testing picture to the assembled database and finding the applicant coordinating highlights. For any item in a picture, intriguing focuses on the picture can be removed to give a "highlight portrayal" of the article. The highlights that are extricated from the preparation picture would then be able to be utilized to distinguish the item. Another significant trait of these highlights is that the relative situations between them in the first scene would not change starting with one picture then onto the next. Filter separates huge number of highlights from the pictures that decreases the mistakes brought about by variations in the normal blunder of all element coordinating mistakes. The highlights that are separated utilizing SIFT is 40, for example 20 for direction focuses and 20 for descriptor focuses. For a picture test, the inclination greatness and direction are figured utilizing pixel contrasts.

Magnitude…

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1), y))}^2 + \sqrt{(L(x, y+1) - L(x, y+1))^2} \quad \ldots\ldots\ldots (7.2)$$

Orientation…

$$\theta(x,y) = \tan^{-1}\left[\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1),y)}\right] \qquad \text{…….(7.3)}$$

For an input image, L(x,y,σ) at the key point's scale σ is taken so that all computations are performed in a scale-invariant manner. For an input image, scale L(x,y) at scale angle, the gradient magnitude $m(x,y)$, and orientation $\theta(x,y)$, are pre-computed using pixel differences: All together, the absolute highlights separated for each character is 79 i.e., 9 - district properties, 30 - corner focuses, 20 – direction focuses and 20 – descriptor focuses. OCR is effectively made utilizing these highlights of the Training Samples.

7.2.5. Classification

11th century written by hand Tamil contents are regularly assembled into four classes to be specific Vowels, Consonants, Composite characters and Aydham. These four classes are taken for characterization reason right now. Customary calculations are far slower than required in light of the fact that the slope-based learning calculation and the parameters must be tuned iteratively. This proposes an algorithm named complex hidden layers extreme learning machine (CEDLM) by Dong(2017). The structure of the CEDLM (select the three-hidden-layer ELM for example) is illustrated in Figure 7.3.

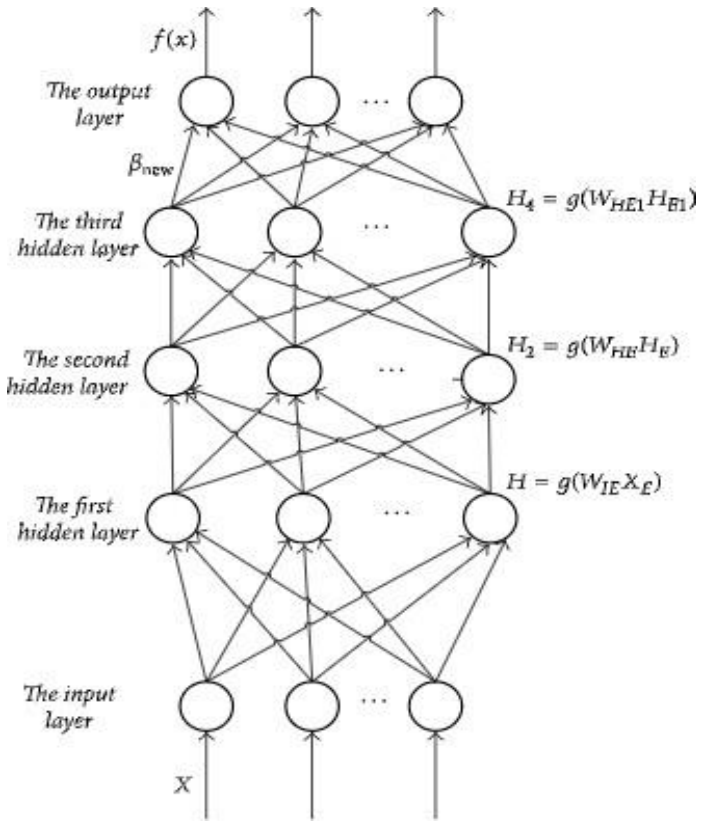


Figure 7.3: Structure of the Three-Hidden-Layer ELM

The output weights matrix β_{new} between the third hidden layer and the output layer is calculated as follows: when the number of hidden layer neurons is less than the number of training samples, β can be expressed as follows:

$$\beta_{new} = \left(\frac{I}{\lambda} + H_4^T H_4\right)^{-1} H_4^T T. \quad \text{........ (7.3)}$$

When the number of hidden layer neurons is more than the number of training samples, β can be expressed as follows:

$$\beta_{new} = H_4^T \left(\frac{I}{\lambda} + H_4 H_4^T\right)^{-1} T. \quad \text{..... (7.4)}$$

The actual output of the three-hidden-layer ELM network can be expressed as follows:

$$f(x) = H_4\beta_{\text{new}}. \quad \text{........ (7.5)}$$

To achieve actual hidden final outcome meets the expected hidden outcome during training process, the operation process is optimized with the network structure parameters starting from the second hidden layer.

The above given is a parameter calculation process of three-hidden-layer EDLM network. Whereas, the purpose of this is to calculate the parameter of the multiple hidden layers ELM network and the final output of the CEDLM network structure Dong(2017).

We can use cycle calculation theory to illustrate the calculating process of the CELDM. If it is found that the hidden layers are increased then the processes can be repeated for further processing.

Algorithm:

Step 1: Assume that the training sample dataset is {X,T}=(x_i,t_i) (i=1,2,3,…..,Q) , where the matrix X is the input sample and the matrix T is the labeled sample. Each hidden layer has l hidden neuron with the activation function g(x)

Step 2: Randomly initialize the weights between W the input layer and the first hidden layer as well as the bias B of the first hidden neurons

$$W_{IE} = [BW], \; X_E = [1X]^T \quad \text{………. (7.6)}$$

Step 3: Calculate the equation $H = g(W_{IE} X_E)$ ……… (7.7)

Step 4: Calculate the weights between the hidden layers and the output layer

$$\beta = (\frac{I}{\lambda} + H^T H)^{-1} H^T T \text{ or } \bar{\beta} = H^T \left(\frac{I}{\lambda} + HH^T \right)^{-1} T \quad \ldots(7.8)$$

Step 5: Calculate the expected output of the second hidden layer $H_1 = T\beta^+$

Step 6: Algorithm steps (4, 5), calculate the weights between the first hidden layer and the second hidden layer and the bias B_1 of the second hidden neurons

$$W_{HE} = g^{-1}(H_1)H_E^+ \quad \ldots\ldots. (7.9)$$

Step 7: Obtain and update the actual output of the second hidden

Layer $H_2 = g(W_{HE}H_E)$ …….. (7.10)

Step 8: Update the weights matrix β between the hidden layer and the output layer

$$\beta_{new} = (\frac{I}{\lambda} + H_2^T H_2)^{-1} H_2{}^T T \quad \text{or} \quad \bar{\beta}_{new} = H_2{}^T \left(\frac{I}{\lambda} + H_2 H_2{}^T \right)^{-1} T \quad \ldots.(7.11)$$

Step 9: If the number of the hidden layer is three, we can calculate the parameters by recycle executing the above operation from step 5 to step 9. Now β_{new} is expressed as follows,

$$\beta_{new} = \beta, H_E = [1H_2]^T \quad \ldots\ldots\ldots (7.12)$$

Step 10: Update the weights matrix β between the hidden layer and the output layer

$$\beta_{new} = (\frac{I}{\lambda} + H_4^T H_4)^{-1} H_4{}^T T \quad \text{or} \quad \bar{\beta}_{new} = H_4^T \left(\frac{I}{\lambda} + H_4 H_4{}^T \right)^{-1} T \quad \ldots\ldots\ldots (7.13)$$

Step 11: If the number N of the hidden layer is more than three and odd number hidden layer, recycle it by executing step 5 to step 9 for (N-1) times. Now β_{new} is expressed as follows

$$\beta_{new} = \beta \, , H_E = \left[(N-2)H_{N-1}\right]^T \quad \ldots\ldots\ldots (7.14)$$

Step 12: If the number N of the hidden layer is more than three and even number hidden layer, recycle it by executing step 5 to step 10 for (N-1) times.

Step 13: Calculate the output $f(x) = H_{N-1}\beta_{new}$ ……… (7.15)

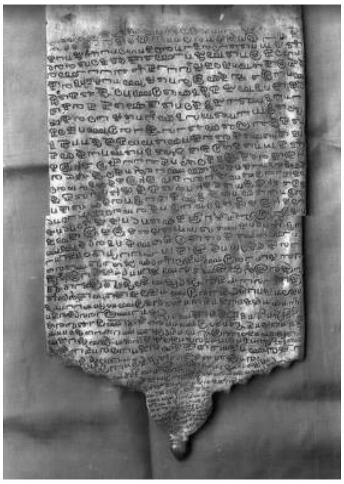
All the H matrix (H_1, H2) must be normalized between the range of −0.9 and 0.9, when the max of the matrix is more than 1 and the min of the matrix is less than −1.

7.3. EXPERIMENTAL RESULTS

To check the productivity and legitimacy of the proposed framework, the framework was tried for precision and the outcomes were contrasted against the aftereffects of past frameworks on similar databases.

So as to ensure the nature of pictures, all pictures were gathered from continuous copper plate pictures taken by us and furthermore from different sources. Preparing sets contained in excess of 70 preparing pictures and in excess of 50 testing tests from CPI-A01 by Olarik (2008). The individual content lines of CPI-A01 database were portioned physically to isolate them into words. Instructional courses were 25 distinctive Tamil words in various sizes, directions, clamor degrees and textual styles. The examinations were led on an AMD Quad-centered, 2 GHz processor, 4 GB DDR3 Ram PC and Windows 8.1 working framework. The code was written in MATLAB language utilizing MATLAB 2011Rb programming. The proposed Copper plate character spotting technique is tried on the pictures of various copper plate engravings gathered

from different places of Tamil Nadu, India. Copper plate engravings made by different period that managed over Tamil Nadu show extraordinary highlights in their style concerning the sort of stones used, cleaning adopted, piece of content used, recording on copper plate with shading, etching the content on copper plate and furthermore dependent on the situation of raising the copper plate in a suitable spot. A considerable lot of these engravings are disintegrated so gravely that it is hard to distinguish the important information, especially when the surface is consumed or carved. Because of hundreds of years of decay, dominant parts of these antiquated writings are in poor condition and numerous content bits are as of now absent. The harm has happened to such a degree, that either the pieces do not exist, or segments are never again conspicuous and the past recuperation happens. The presentation aftereffect of the character spotting process on such pictures is likewise announced right now.



.

‘

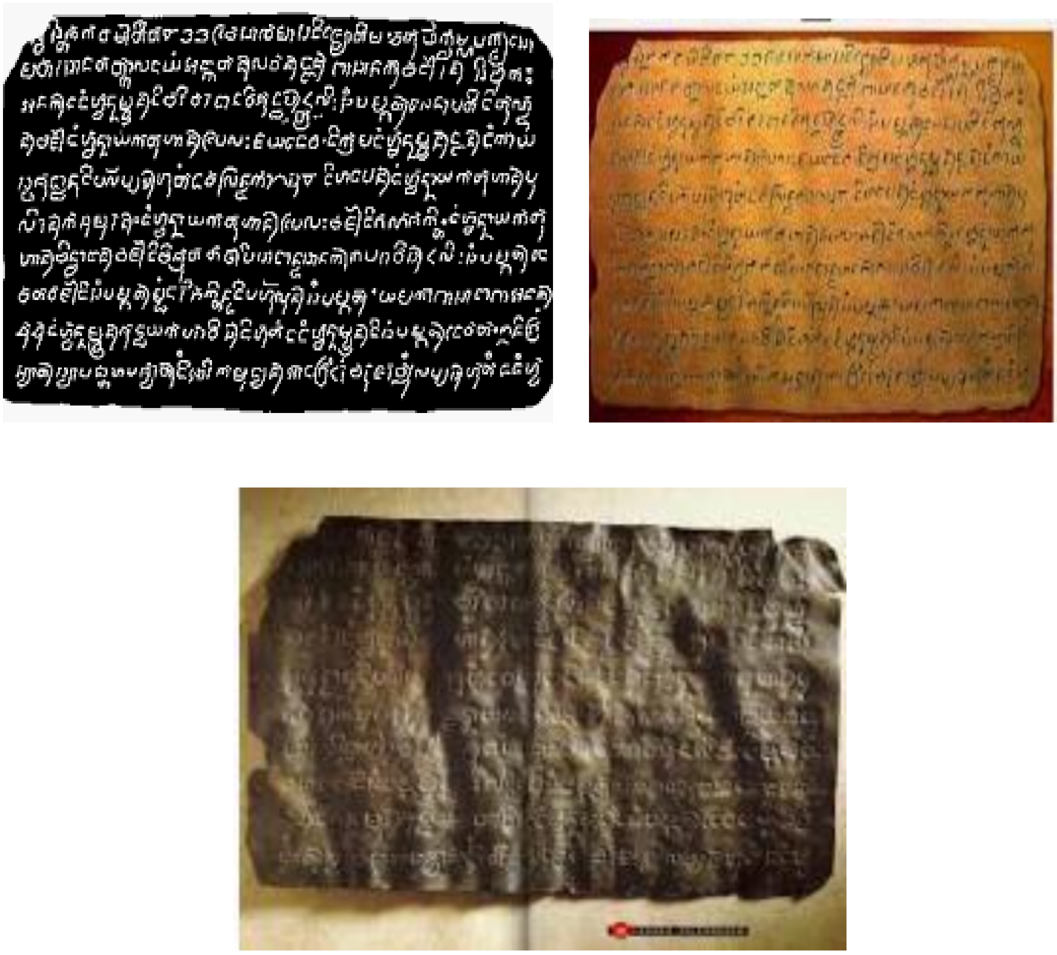


Figure. 7.4: Sample Copper Plate images

At first, the element dataset contains full 14 highlights, the preparation dataset is of measurement 102x14 and the testing dataset is of measurement 22x14. The two sets are applied to EDLM and CEDLM. The average classification accuracy achieved in the test data is taken as the classification performance criteria of the problem. Figure 7.2 depicts the average testing classification correct percentage for the ELM, EDLM and CEDLM algorithm is shown clearly. Regression Problems. To test the performance of the regression problems, several widely used functions are listed as stated by Liang et al (2006).We use these functions to generate a dataset which includes random selection of

sufficient training samples and the remaining is used as a testing samples, and the activation function is selected as the hyperbolic tangent function $g(x) = (1 - e{-}x)/(1 + e{-}x)$.

(1) $f_1(x) = \sum_{i=1}^{D} x_i.$

(2) $f_2(x) = \sum_{i=1}^{D-1}(100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2).$

(3) $f_3(x) = -20e^{-0.2\sqrt{\sum_{i=1}^{n} x_i^2/D}} - e^{\sum_{i=1}^{n}\cos(2\pi x_i)/D} + 20.$

The symbol D that is set as a positive integer represents the dimensions of the function we are using. The function $f_2(x)$ and $f_3(x)$ are the complex multimodal function.

Table 7.1: RMSE Values CPI-A01 Dataset

S.No	**Algorithm**	**Training RMSE**	**Testing RMSE**
1	Extreme Learning Machine (ELM)	$8.6233E{-}9$	$1.0797E{-}8$
2	Extreme Deep Learning Machine (EDLM)	$1.8428E{-}6$	$1.5204E{-}5$
3	Complex Extreme Deep Learning Machine (CEDLM)-Proposed System		
	$f_1(x)$	$8.722E{-}15$	$1.3055E{-}14$
	$f_2(x)$	0.0011	0.0019
	$f_3(x)$	0.2110	0.4177

Table 7.2: Comparison between the Proposed System Classifier and Different Classifiers on the CPI −A01 Database

S. No	System Classifier	Max Test Accuracy
1	Probabilistic Neural Network (PNN)	70.18%
2	Support Vector Machine (SVM)	72.4%
3	Extreme Learning Machine (ELM)	78.13%
4	Extreme Deep Learning Machine (EDLM)	85.87%
5	Complex Extreme Deep Learning Machine (CEDLM) -Proposed System	92.05%

In addition to this, the classifier framework in Rafael (2007) relies upon some heuristic punishments and division procedures that incredibly influence the Sift descriptor exactness, yet our framework is free of any component descriptors and it utilizes an amazing arrangement of interpretation and scale invariant highlights.

7.3 SUMMARY

Copper Plate Optical Character Recognition (CPOCR) for composed content is an extreme testing and open territory of research. This work builds a copper plate Tamil OCR for manually written words that is dependent on a blend of the Complex Extreme Deep Learning Machine (CEDLM) classifier with Multi hidden Layer Feed Forward Network and fact-based component determination. Toward the start, the framework

utilized a 14 highlights dataset. At that point, information was sustained into EDLM organizer, which is a quick and basic multi concealed layer feed forward system (MLFN). The framework accomplished high acknowledgment precision of 92.05% for various examples in an extremely brief timeframe. The CEDLM calculation acquires the qualities of customary ELM that arbitrarily introduces the loads and predisposition (between the information layer and the main covered up layer), which also embraces a piece of the ELM calculation and utilizes the backward initiation capacity to compute the loads and inclination of concealed layers (with the exception of the primary shrouded layer). At this point, we make the real concealed layer yield exactly to the normal shrouded layer yield and utilize the parameters that got above to figure the real yield. In the capacity relapse issues, this calculation decreases the least mean square blunder. In the datasets order issues, the normal precision of the various arrangements is fundamentally higher than that of the ELM and EDLM organize structure.

CHAPTER 8

8. EXPERIMENTAL RESULTS ANALYSIS AND DISCUSSION

8.1 INTRODUCTION

In this research work, a prototype system is developed which converts the copper plate handwritten characters into understandable and standard format of Tamil text for seamless conversion of handwritten document into typed documents with ease. This thesis presents a complete Optical Character Recognition (OCR) technique followed by Handwritten to Typed Text conversion. Various algorithms for optical character recognition have been studied and analyzed. Based on the analysis, a new algorithm was developed and implemented in this work to make the system provide better result. The advantage of this prototype is that it performs the handwritten character recognition of Tamil language in a single system. This work is one of its kind of implementation exclusively done for Tamil character recognition.

This research study proposes Classification algorithms namely Extreme Deep Learning Machine (EDLM) and Complex Extreme Deep Learning Machine (CEDLM) and compares them with existing classification techniques like Probabilistic Neural Network (PNN), Support Vector Machine (SVM) and Extreme Deep Learning Machine (EDLM) for the collected copper plate Tamil character written images. The maximum test accuracy achieved in this study is listed in the below Table 8.1. It also documents the promising results shown by the proposed classification algorithm when compared to existing algorithms.

Table 8.1: Max Test Accuracy for Different System Classifier and Proposed Classifier

S.NO	System Classifier	Max Test Accuracy
1	Probabilistic Neural Network (PNN)	70.18%
2	Support Vector Machine (SVM)	72.4%
3	Extreme Learning Machine (ELM)	78.13%
4	Extreme Deep Learning Machine (EDLM)	85.87%
5	Complex Extreme Deep Learning Machine (CEDLM) Proposed System	**92.05%**

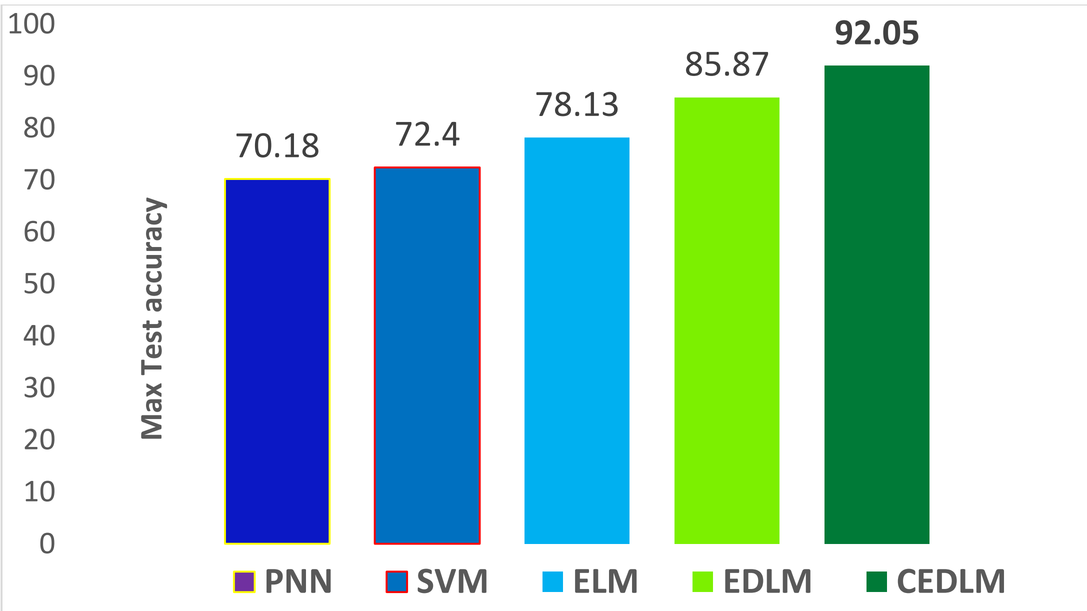


Figure 8.1: Comparison Chart of Max Test Accuracy for Different System Classifier and Proposed Classifier

In this work, the proposed non-toxic chemical treated copper plate images are taken to improve the system performance. Some sample copper plate images are taken for testing, on which the solution is applied. To introduce an eco-friendly phytochemical technique for the removal of corrosion mechanism from copper objects victimization, Bryophyllum calcynium (Ranakalli plant) is used as the main course material for the corrosion removal in conjunction with supplementary binding agents. The residues of this process are bio-degradable and hence do not harm the environment. Moreover, the corrosion effect on the copper plates on using this composition is comparatively lesser with respect to other chemical treatments that are in practice currently.

Firstly 100 grams of Bryophyllum calcynium leaves and Clean Bryophyllum calcynium leaves with water are used. Secondly, 100 grams of raw rice, 5 grams of fenugreek seeds and 25 grams of Split Black gram are soaked for 3 hours in water. Then the water is drained and Bryophyllum calcynium leaves are grinded into a paste. The resultant paste is allowed to settle for about 6 hours for the Fermentation process to take place, without adding any external chemical agents.

The fermented paste is applied over the corroded metallic copper plate which is left with the binding agent for an hour. The resultant plate is then washed with water to remove the binding agent.

It was observed that the results obtained after applying the above process is more promising than before. The maximum test accuracy achieved is listed in the below Table 8.2.

Table 8.2: Max Test Accuracy for Proposed System Classifier before and after Applying Solution

S. NO	System Classifier	Before applying Solution	After applying Solution
		Max Test Accuracy	
1	Extreme Learning Machine (ELM)	72.68%	78.13%
2	Extreme Deep Learning Machine (EDLM)	79.05%	85.87%
3	Complex Extreme Deep Learning Machine (CEDLM) Proposed System	86.55%	92.05%

On the whole, the result accuracy obtained when using CEDLM simulations technique towards recognizing Tamil characters on antiquated copper plates by Eco-friendly method is higher when compared to other prevailing techniques along with traditional ELM method. Figure 8.2 shows the comparison chart for available ELM and proposed EDLM and CEDLM.

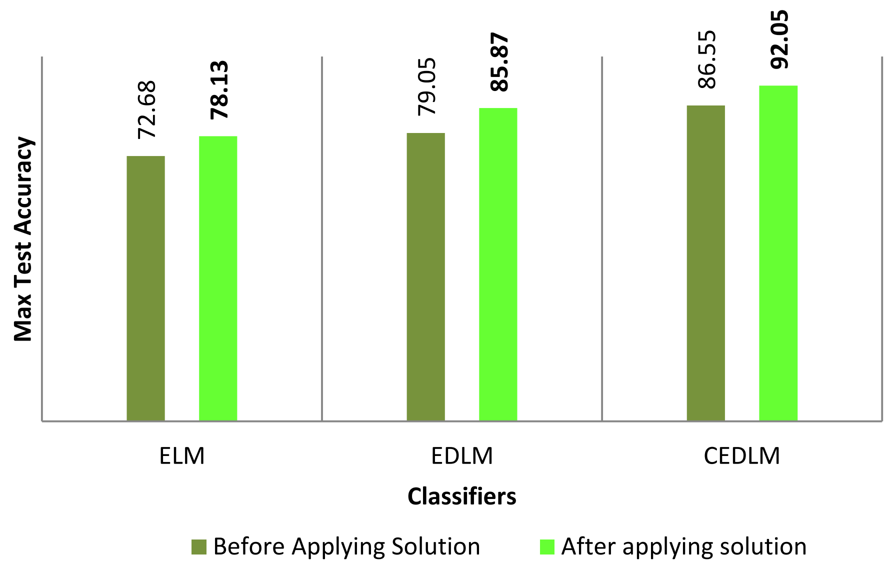


Figure 8.2: Comparison Chart of Max Test Accuracy for Proposed System Classifier before and after Applying Solution

8.2 SUMMARY

The framework accomplished high recognition precision of 92.05% for various examples tried. The CEDLM calculation acquires the qualities of customary ELM that arbitrarily introduces the loads and predisposition (between the information layer and the main covered up layer), which also holds a portion of the ELM calculation and utilizes the backward initiation capacity to compute the loads and inclination of concealed layers (with the exception of the primary shrouded layer). In the datasets used, the normal precision of various arrangements is fundamentally higher than that of the ELM and EDLM organized structures, the result of CEDLM supports in elevating the existing system structure.

Hence this system focuses on developing libraries for few Tamil characters. Presently, 30 character-sets are trained. The feature work will be focused on training more Tamil characters and accuracy of the system to be enhanced. The system could be integrated with other digital systems to keep it as simple as possible and also economical, along with operational feasibility in housekeeping the digital contents.

CHAPTER 9

9. CONCLUSION AND FUTURE SCOPE

9.1 CONCLUSION

In this chapter, the main contributions delivered and the significant achievement acquired from this research work is summarized. The conclusion, which follows the summary, highlights the research contributions delivered in the field of copper plate Tamil character recognition using image processing. Moreover, on the view of providing the future exploring possibilities to researches that follow, the present limitations and expansion possibilities of this system are also briefed.

Copper Plate Optical Character Recognition (CPOCR) for composed content is at present an open territory of research. The main aspect of this thesis is to concentrate on Tamil character recognition using copper plate Tamil text image restoration process. The present constraints in these fields of character recognition are either related to feature extraction or classification difficulties. This thesis is focused on overcoming these difficulties faced on feature extraction and classification as both of them have equally important roles to play in character recognition.

Tamil scripts are normally grouped into four classes namely Vowels, Consonants, Composite characters and Aydham. These four classes are taken for classification purpose in this research work. Traditional algorithms are far slower than required

because of their gradient based learning algorithm and the parameters have to be tuned iteratively. And therefore, Extreme Learning Machine (ELM) is used for classification.

The performance of ELM is compared with Probabilistic Neural Network (PNN) and it is observed that 70.19% and 78.73% of accuracy is attained by PNN and ELM respectively. In order to increase the accuracy of classification further, Extreme Deep Learning Machine (EDLM) and Complex Deep ELM (CEDLM) are used. Extension of an ELM from real domain to complex domain is known as Complex ELM. The performance of EDLM and CEDLM is measured by comparing it with ELM. After applying eco-friendly cleansing process, our proposed algorithms EDLM and CEDLM give the highest rate of performance measures of 85.87% and 92.05% when compared to ELM.

The test accuracy attains its maximum of 92.05% result which is better when compared to the results of the existing classification techniques like PNN, SVM and EDLM for collected copper plate Tamil character written images. In such cases, the CEDLM can improve the presentation of the system structure. Therefore, this proposal has increased the value for the field of Tamil Character Recognition precisely in the field of copper plate recognition.

This is extremely useful for researchers who are engaged in recognizing the metallic inscriptions worldwide as the same kind of metals can be found in most of the scripts used in the world.

9.2 FUTURE SCOPE OF RESEARCH

There is always a way better than the one that has been followed. Every versatile solution will have adequate flexibility for further extension. In any case, there are difficulties related with transcribed Tamil character acknowledgment, which has huge degree of scope for future research.

The composite characters have the essential structure looking like the consonants with minor alteration on the fundamental structure or have a supporting character which misleads to other characters.

Segmentation of content from non-content foundation is unexplored (for all the dialects) and has an incredible research potential.

Furthermore, many productive designs could be created for execution of Tamil character recognition. Using the same technique, recovery of characters is possible globally over any non-headline-based scripts.

In many places, it is prohibited even to take photocopy of the copper plates without incorporating eco-friendly cleaning processes. So, the data sample size was reduced to what was available. This research is a promising step towards bringing back vital information from even partially deteriorated copper plates which is otherwise left unattended. If this research is properly extended, which can be of use to research departments of archaeology and epigraphy, many precious copper plates information can be extracted and preserved for our future references.

ABSTRACT

Refurbishing schemes towards restoring and protecting classical metal-plates with inlaid inscriptions and/or carvings reflect artistic impressions and useful information of our history. They enable us to comprehend individual traits, regarding various phases of time possessing different propensities and customs. Every antiquated composition is available in different forms of structures like stones, carvings in metal-plates, inscriptions written in stacks of palm-leaf and original copies of documented papers. As such, hundreds and thousands of authenticated copper-specific ancient landmarks, significant data relating the history and culture of that era. However, these copper-based antiques have been gradually deteriorating and crumbling as a result of ecological contaminations and other organic and related anthropogenic exercises. Therefore, there is an imminent need to preserve these copper landmarks and to re-establish (refurbish) them for our future use and references. Hence, *Copper Plate Optical Character Recognition* (CPOCR) for the included contents is an open territory for research and testing. Relevantly, this thesis explores the underlying, traditional aspects of data acquisition in copper plates and identifies customary interpretation techniques. With regard to such efforts, an interactive tool is proposed here for epigraphist compatibility towards rescuing valuable inscriptions of the past so as to identify the tarnished characters inscribed in copper-plates. The

proposed technique targets retrieving valuable information from copper-plates in an efficient and reliable method with minimal processing time. Objectively, the proposed scheme in this thesis would help viably identifying and transcribing the smudged and/or problematic inscriptions/carvings seen in copper- plates of Tamil writings/images; and, relevant restorations can be undertaken thereof in character recognition that forms the fundamental aspect of Tamil text image- processing strategies. Relevant problems in character recognition is generally allied either to feature extraction or in classifications. Jointly, extraction and classification play important roles and this thesis focuses on both issues equally. Tamil scripts are normally grouped under four major heads namely Vowels, Consonants, Composite characters and Aydham (or *ahk* - the special letter "ஃ"). The aforesaid four classes are considered for classification purposes in this research work. Traditional algorithms in pertinent machine-learning (ML) efforts are far slower in as much as the associated gradient-based learning algorithm and the underlying parameters have to be tuned iteratively. Therefore, a need for an Extreme - version Learning Machine (ELM) is suggested and adopted here for classification.

The performance of such ELM is compared with Probabilistic Neural Network (PNN); and, the study shows that about 70.19%, 78.73% of accuracy is attained with PNN and ELM respectively. Further, to improve

the accuracy of classification a Complex ELM method is indicated. Proposed extension refers to using the ELM in a domain of complex field (termed as Complex ELM) and the performance of Complex ELM is ascertained by comparing it against the traditional ELM. It is observed that Complex ELM (CELM) yield better rate of classification accuracy of b about 80.30%. This thesis describes a Copper Plate Optical Character Recognition scheme for manually written Tamil words. The suggested method uses Complex and Extreme Deep Learning Machines (CEDLM) along with a Multi - Hidden Layer Feed-forward Network and use of factual or real data based component-determination approach is pursued. In the initial analysis, the test framework uses 14 datasets and subsequently, the information is adopted in the EDLM and basic multi-hidden layer feed-forward system (MLFN). The test framework enables high recognition precision (to an extent of about 92.05%) in various test ensembles involving brief time frames of execution. This research is limited to data recovery from 11th century copper plates containing Tamil inscriptions alone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
CPCR	Copper Plate Character Recognition
CV	Consonant Vowel
GA	Genetic Algorithm
ELM	Extreme Learning Machine
EDLM	Extreme Deep Learning Machine
MLFN	Multi-shrouded Layer Feed-Forward Networks
OCR	Optical Character Recognition
PSO	Particle Swarm Optimization
CELM	Complex Extreme Learning Machine
PNN	Probabilistic Neural Network
DE	Differential Evolution
HMM	Hidden Markov Models
SOM	Self-Organizing Map
MLP	Multi-Layer Perceptron
SVM	Support Vector Machine
DTW	Dynamic Time Traveling
SHMM	Structural Hidden Markov Model
EDTA	Ethylene Diamine Tetra Acetate
PPV	Positive Predictive Value
NPV	Negative Predictive Value
TP	True Positives
FP	False Positives

TN	True Negatives
FN	False Positives
SIFT	Scale Invariant Feature Transform
RMSE	Root Mean Square Error
CPI	Characters Per Inch
CSR	Character Spotting Result

SIMULATION TECHNIQUES TOWARDS RECOGNIZING TAMIL CHARACTERS INSCRIBED IN ANTIQUATED COPPER-PLATES: APPLICATION OF COMPLEX DEEP-LEARNING MACHINE ON ECO- FRIENDLY CLEANSED PLATES

A Thesis

Submitted by

R. INDRA GANDHI

[Reg. No. D15CS512]

In partial fulfilment for the award of the degree

of

DOCTOR OF PHILOSOPHY

Under the Supervision of

Dr. M. PONNAVAIKKO

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF SCIENCE AND HU ENGINEERING

BHARATH INSTITUTE OF HIGHER EDUCATION AND RESEARCH

CHENNAI -600 073

JANUARY 2021